Applied Probability II

Section 2: Recap on previous material

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 2: Recap on previous material

Section 2.1: Applied Probability I

Topics covered

- Random variables: discrete and continuous
- Discrete distributions
- Continuous distributions

Random variables: discrete and continuous

- What is a random variable?
  - the ‘outcome’ from an ‘experiment’
- You won’t know the outcome in advance, but you can think about:
  - what are the possible values or range of values?
  - with what probability are certain values, or ranges of values, likely to occur?
- Random variables can be discrete or continuous.
- We talk about the ‘distribution’ of a random variable.

Discrete distributions

- To characterise a discrete distribution, we can think about:
  - possible values
  - probability mass function (pmf), $P(X = x)$
  - cumulative distribution function (cdf), $P(X \leq x)$
  - mean or expected value, $\mathrm{E}[X]$
  - variance, $\mathrm{Var}(X)$

- Side note: when do we use capital X and when do we use little x?

- Some commonly known discrete distributions
  - Bernoulli
  - Binomial
  - Poisson
  - Geometric
  - Negative binomial
  - Hypergeometric
  - Uniform

To help think about discrete distributions here is a link to some music from Dr Rafael de Andrade Moral, a talented collaborator of mine! https://youtu.be/ZINXFoQMZVs

Discrete distributions contd.

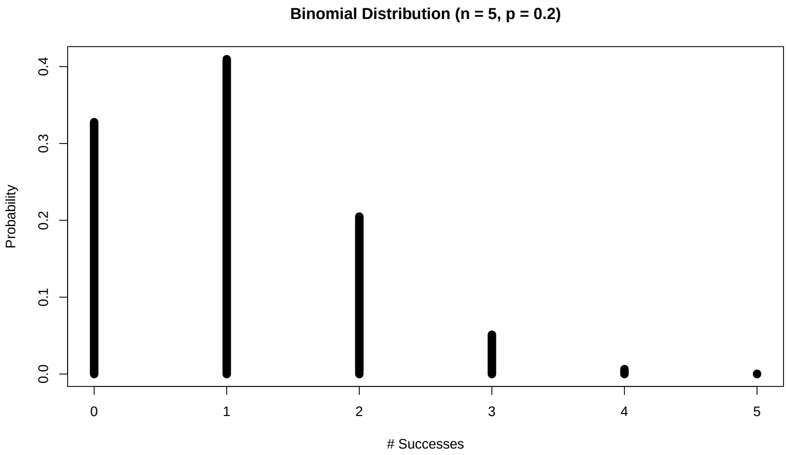
For the binomial distribution, consider $X \sim \text{Bin}(n = 5, p = 0.2)$. The random variable X is the number of 'successes' in $n = 5$ trials.

How might we characterise this distribution?

- possible values: $\{0, 1, 2, 3, 4, 5\}$
- probability mass function (pmf) $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{5}{k} 0.2^k (0.8)^{5-k}$
- cumulative distribution function (cdf), $P(X \leq x)$
- mean or expected value, $\text{E}[X] = np = 5 \times 0.2 = 1$
- variance, $\text{Var}(X) = npq = 5 \times 0.2 \times 0.8 = 0.8$

Discrete distributions contd.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{5}{k} 0.2^k (0.8)^{5-k}$$



Continuous distributions

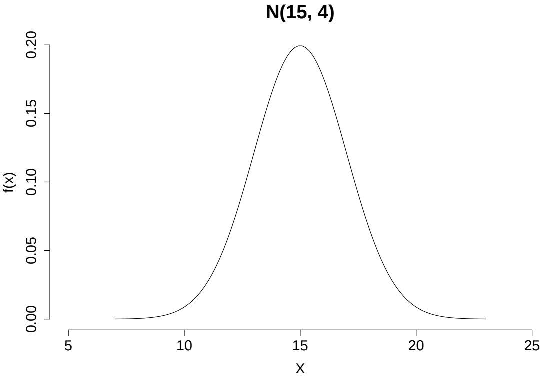
- To characterise a continuous distribution, we can think about:
  - the range of values
  - probability density function (pdf), $f(x) \geq 0$, and $\int_{-\infty}^{\infty} f(x)dx = 1$
  - cumulative distribution function (cdf), $F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$.
- mean or expected value, $\mathrm{E}[X]$
- variance, $\mathrm{Var}(X)$

- Some commonly known continuous distributions
  - Normal
  - Uniform
  - Exponential

Continuous distributions contd.

The normal (or Gaussian) distribution, consider $X \sim \mathrm{N}(\mu = 15, \sigma^2 = 4)$.

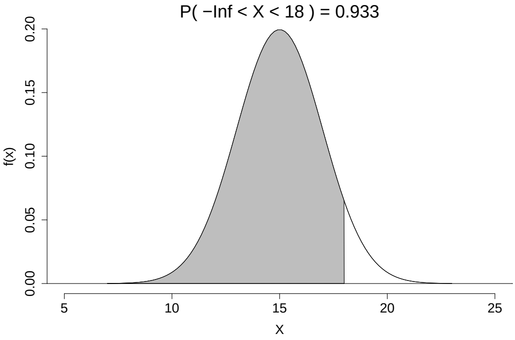
Probability density function (pdf): $\int_{-\infty}^{\infty} f(x)dx = 1$



Continuous distributions contd.

The normal (or Gaussian) distribution, consider $X \sim \mathrm{N}(\mu = 15, \sigma^2 = 4)$.

Cumulative distribution function (cdf), $F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$.

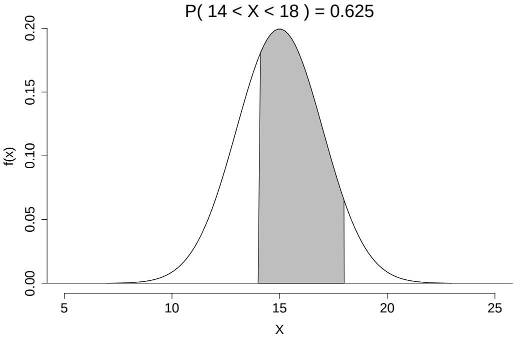


$$F(18) = P(X \leq 18) = \int_{-\infty}^{18} f(x)dx$$

Continuous distributions contd.

The normal (or Gaussian) distribution, consider $X \sim \mathrm{N}(\mu = 15, \sigma^2 = 4)$.

Cumulative distribution function (cdf), $F(a) = P(a \leq X \leq b) = \int_a^b f(x)dx$.



$$F(18) = P(14 \leq X \leq 18) = \int_{14}^{18} f(x)dx$$

Other topics

- Joint distributions
  - discrete
  - continuous
  - marginal distributions
  - independence
  - expectation
  - moment generating functions
  - covariance and correlation
  - conditional distributions
  - Bayes' rule
  - Bivariate normal distribution
- Law of Large Numbers
- Monte-Carlo Simulation
- Introduction to regression analysis

Section 2.2: Populations versus samples

Populations

Examples include

- All adults in Ireland
- People at risk of a particular disease
- All trees of a particular species

What population ‘parameters’ might we be interested in?

Samples

Consider the population:

- adults in Ireland

and population parameter of interest:

- the mean height.

How might we take random samples from this population?

- wrong answers ("bad" samples!) first!
- and good samples?

Terminology

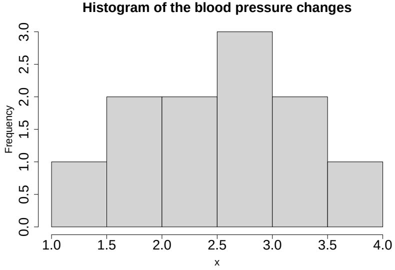
- **Parameter**: Population characteristic.
  - For example, μ, σ, σ^2, π.
- **Sample statistic**: Any quantity computed from values in a sample.
  - For example $\bar{x}$, s, s^2, p.
- But what about $\bar{X}$, S, S^2, P? What do these represent?
- **Statistical inference**?

Section 2.3: Hypothesis testing and confidence intervals

Pharmaceutical example

A pharmaceutical company has developed a drug they believe will help cure a particular disease and has acquired legal permission to test the drug on humans. However, there are safety concerns that the drug will have an adverse effect on blood pressure by increasing it. They run an experiment to estimate the average change in blood pressure for patients who take the new drug. They only have permission to test the drug on eleven patients.

The eleven recordings are: 1.1, 1.8, 2, 2.4, 2.5, 2.8, 2.9, 3, 3.4, 3.4, 4. The mean, $\bar{x} = 2.66$, standard deviation, $s = 0.824$ and the number of observations, $n = 11$.



Pharmaceutical example: confidence interval

Population parameter of interest: μ, the true mean change in blood pressure for the population of people who take the drug.

Suppose the data represent a random sample from the population of all people who take the drug. Then we would estimate the population mean μ by the sample mean $\bar{x}$, i.e. we **estimate** the mean change in blood pressure to be $\bar{x} = 2.66$.

95% confidence interval for μ: (2.11, 3.22)

This is computed as

$$\bar{x} \pm t_{\nu, \frac{0.05}{2}} \frac{s}{\sqrt{n}}$$

Let's do a Poll. How should we interpret the confidence interval?

A. We are 95% sure the sample mean lies in the interval.

B. We know that 95% of the observations studied lie in the interval.

C. If the experiment is repeated, we are 95% sure the next sample mean will lie in the interval.

Pharmaceutical example: confidence interval

Here is our confidence interval again:

95% CI for μ: (2.11, 3.22)

How is this confidence interval interpreted?

Let's answer these questions first:

- Would we get the same CI if we repeated the experiment?
- What do we know about $\bar{x}$ and our 95% confidence interval?
- What do we know about μ and our 95% confidence interval?

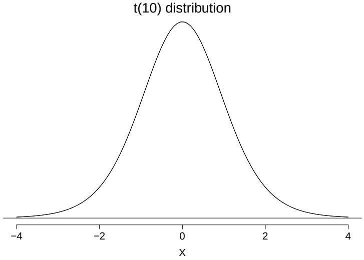
We are 95% confident that our population parameter of interest, μ, lies in this interval.

What would we expect to happen if we repeated the experiment 100 times?

Pharmaceutical example hypothesis test

- $H_0 : \mu = 0$ vs $H_A : \mu > 0$.
- The observed test statistic is $T_{obs} = \frac{\bar{x}-0}{s/\sqrt{n}} = 10.72$
- P-value $= \mathrm{P}(t_{10} \geq 10.72) < 0.001$.
- Conclusion: We reject H_0 that $\mu = 0$ at $\alpha = 0.05$. We have evidence that the population average blood pressure change is greater than 0, i.e., evidence that the drug increases blood pressure.

Note about p-values. The p-value is the probability of observing a test statistic as extreme or more extreme than what was observed, given that the null hypothesis is true. Here, we assume that the distribution of the test statistic under the null hypothesis is:



How do confidence intervals and hypothesis tests work?

In this module, we will look at the underlying theory behind CIs and hypothesis tests.

Some closing thoughts:

- What assumptions did we make for the confidence interval for the pharmaceutical example?
- Are those assumptions reasonable?
- What assumptions did we make for the hypothesis test?