Introductory Statistics

Tatiana Tyukina
tt51@leicester.ac.uk

Topic 0 - Introduction. Part 2

- Topic 0: Introduction
- Topic 1: Descriptive Statistics
- Topic 2: Estimators: Point estimation, Interval estimation
- Topic 3: Hypothesis Testing
- Topic 4: Goodness of Fit
- Topic 5: Bayesian Estimation.

Introduction

- What is statistics?
- Examples
- How do we use statistics?
- Population vs Sample
- Random variable, Probability Distributions
- Moments and Moment Generating Functions

DISCRETE PROBABILITY DISTRIBUTIONS

- A random variable X is said to be *uniformly distributed* over the numbers $1, 2, 3, \ldots, n$ if
$$P(X = i) = \frac{1}{n}, \text{ for } i = 1, 2, \ldots, n.$$
The *uniform probability function* is defined as:
$$p(i) = P(X = i) = \frac{1}{n}, \text{ for } i = 1, 2, \ldots, n.$$
The *cumulative uniform distribution function* is:
$$F(x) = P(X \leq x) = \sum_{i \leq x} p(i) = \sum_{i \leq x} \frac{1}{n} = \frac{x}{n},$$
for $0 \leq x \leq n$.

Discrete Probability Distributions

- *Bernoulli experiment and trials*
  - $S = \{'\text{success}', '\text{failure}'\}$, i.e. $X \in \{1, 0\}$;
  - *Bernoulli trials* are the fixed sequence of n identical repetitions of the same Bernoulli experiment;
  - for each trial probability of success is p, probability of failure is $q = 1 - p$, these probabilities are constant for every trial;
  - the outcome of each experiment is independent on previous experiments and does not influence any subsequent outcomes.

Discrete Probability Distributions

- *Bernoulli experiment and trials*
- a *binomial random variable* is a discrete random variable that describes the number of successes ($X = 1$) in a sequence of n Bernoulli trials. The *binomial probability function* is defined as:

$$p(k) = P(X = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}, \; k = 0, 1, 2, 3..., n.$$

Cumulative binomial distribution function is:

$$F(x) = P(X \le x) = \sum_{k \le x} p(k) = \sum_{k \le x} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k},$$

for $0 \le x \le n$.

DISCRETE PROBABILITY DISTRIBUTIONS

- *Bernoulli experiment and trials*
- a *binomial random variable* is a discrete random variable that describes the number of successes ($X = 1$) in a sequence of n Bernoulli trials. The *binomial probability function* is defined as:

$$p(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ k = 0, 1, 2, 3\ldots, n.$$

Cumulative binomial distribution function is:

$$F(x) = P(X \leq x) = \sum_{k \leq x} p(k) = \sum_{k \leq x} \binom{n}{k} p^k (1-p)^{n-k},$$

for $0 \leq x \leq n$.

- Hypergeometric random variable

Discrete Probability Distributions

- The *geometric distribution* is either one of two discrete probability distributions:

Discrete Probability Distributions

- The *geometric distribution* is either one of two discrete probability distributions:
  - The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $S = \{1, 2, 3, ...\}$;

DISCRETE PROBABILITY DISTRIBUTIONS

- The *geometric distribution* is either one of two discrete probability distributions:
  - The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $S = \{1, 2, 3, ...\}$;
  - The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set $S = \{0, 1, 2, , ...\}$.

DISCRETE PROBABILITY DISTRIBUTIONS

- The *geometric distribution* is either one of two discrete probability distributions:
  - The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $S = \{1, 2, 3, ...\}$;
  - The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set $S = \{0, 1, 2, , ...\}$.

  The corresponding probability distribution functions and the cumulative distribution functions are:
  - for X representing number of 'successes'

$$p(k) = P(X = k) = (1-p)^{k-1}p, \text{ for } k = 1, 2, 3, ...$$

$$F(x) = P(X \leq x) = \sum_{k \leq x} (1-p)^{k-1}p = 1 - (1-p)^k$$

DISCRETE PROBABILITY DISTRIBUTIONS

- The *geometric distribution* is either one of two discrete probability distributions:
  - The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $S = \{1, 2, 3, ...\}$;
  - The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set $S = \{0, 1, 2, , ...\}$.

  The corresponding probability distribution functions and the cumulative distribution functions are:
  - for X representing number of 'successes'

$$p(k) = P(X = k) = (1-p)^{k-1}p, \text{ for } k = 1, 2, 3, ...$$

$$F(x) = P(X \leq x) = \sum_{k \leq x} (1-p)^{k-1} p = 1 - (1-p)^k$$

  - for Y representing number of 'failures'

$$p(k) = P(Y = k) = (1-p)^k p, \text{ for } k = 0, 1, 2, 3, ...$$

$$F(y) = P(Y \leq y) = \sum_{k \leq y} (1-p)^k p = 1 - (1-p)^{k+1}$$

DISCRETE PROBABILITY DISTRIBUTIONS

- A *Poisson random variable* describes the random events occurring in continuous fixed units of time and space.
  - k is the number of times an event occurs in an interval and $k = 0, 1, 2, ...$
  - The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
  - The average rate at which events occur is independent of any occurrences.
  - Two events cannot occur at exactly the same instant; instead, at each very small sub-interval, either exactly one event occurs, or no event occurs.

DISCRETE PROBABILITY DISTRIBUTIONS

- A *Poisson random variable* describes the random events occurring in continuous fixed units of time and space.
  A discrete random variable X has a *Poisson distribution* with parameter $\lambda > 0$ if it's probability mass function can be described by:

$$p(k) = P(X = k) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!}, \text{ for } k = 0, 1, 2, \ldots$$

  The cumulative distribution function is:

$$F(x) = P(X \leq x) = \sum_{k \leq x} p(k) = \sum_{k \leq x} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \cdot \sum_{k \leq x} \frac{\lambda^k}{k!},$$

  for $0 \leq x < \infty$.

Continuous Probability Distributions

- The *continuous uniform distribution* describes an experiment where there is an arbitrary outcome that lies between certain bounds. The random variable X is defined on the closed interval $[a, b]$ (or open interval (a, b)).

Continuous Probability Distributions

- The *continuous uniform distribution* describes an experiment where there is an arbitrary outcome that lies between certain bounds. The random variable X is defined on the closed interval $[a, b]$ (or open interval (a, b)). The probability density function is defined as:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

The cumulative distribution function is:

$$F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases}$$

for $0 \leq x < \infty$.

Continuous Probability Distributions

- A random variable X is called *Gaussian* or *normal* if its probability density function $f(x)$ is of the form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where $-\infty < x < \infty$ and μ and σ are two parameters, such that $-\infty < \mu < \infty$, $\sigma > 0$.
The cumulative distribution function is:

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

for $-\infty < x < \infty$.

CONTINUOUS PROBABILITY DISTRIBUTIONS

- A random variable X is called *Gaussian* or *normal* if its probability density function $f(x)$ is of the form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where $-\infty < x < \infty$ and μ and σ are two parameters, such that $-\infty < \mu < \infty$, $\sigma > 0$.
The cumulative distribution function is:

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

for $-\infty < x < \infty$.

CONTINUOUS PROBABILITY DISTRIBUTIONS

- The **exponential distribution** is the probability distribution of the time between events in a *Poisson point process*, i.e., a process in which events occur continuously and independently at a constant average rate.
  The probability density function of an exponentially distributed random variable X ($x \in [0, \infty)$) can be parametrised in two ways:

by *a rate parameter* $\lambda > 0$	by *a scale parameter* $\beta > 0, \beta = 1/\lambda$
the pdf has the form: $f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0, & x < 0 \end{cases}$;	the pdf has the form: $f(x) = \begin{cases} \frac{e^{-x/\beta}}{\beta} & x \geq 0 \\ 0, & x < 0 \end{cases}$;
the cumulative distribution function is: $F(x) = \begin{cases} (1 - e^{-\lambda x}) & x \geq 0 \\ 0, & x < 0 \end{cases}$	the cumulative distribution function is: $F(x) = \begin{cases} (1 - e^{-x/\beta}) & x \geq 0 \\ 0, & x < 0 \end{cases}$

Probability Distributions in R

- Discrete probability distributions

Distribution	pdf	cdf	inverse cdf	random deviates
Binomial	dbinom	pbinom	qbinom	rbinom
Geometric	dgeom	pgeom	qgeom	rgeom
Hypergeometric	dhyper	phyper	qhyper	rhyper
Poisson	dpois	ppois	qpois	rpois

PROBABILITY DISTRIBUTIONS IN R

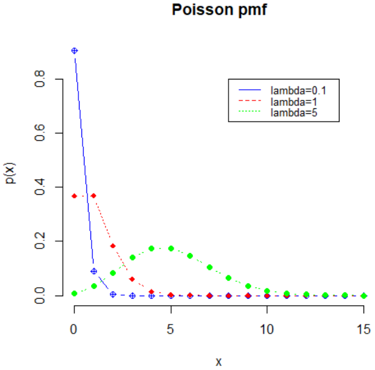
- Discrete probability distributions

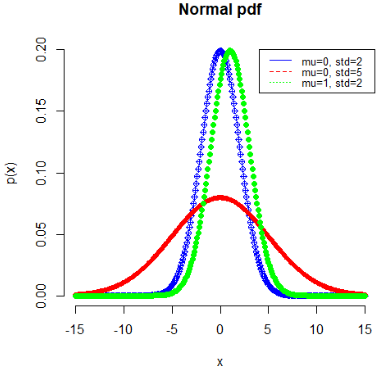
Distribution	pdf	cdf	inverse cdf	random deviates
Binomial	dbinom	pbinom	qbinom	rbinom
Geometric	dgeom	pgeom	qgeom	rgeom
Hypergeometric	dhyper	phyper	qhyper	rhyper
Poisson	dpois	ppois	qpois	rpois

- Continuous probability distributions

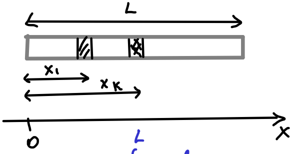
Distribution	pdf	cdf	inverse cdf	random deviates
Normal	dbinom	pbinom	qbinom	rbinom
Exponential	dexp	pexp	qexp	rexp
Uniform	dunif	punif	qunif	runif

Probability Distributions in R





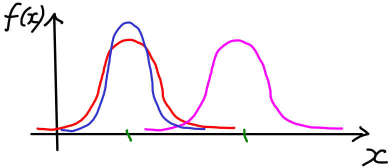
Moments



$$x_{cm} = \frac{\int_0^L x \cdot dm}{\int_0^L dm}$$

$$\rho = \frac{dm}{dx} \text{ - density}$$

$$x_{cm} = \frac{\int_0^L x \rho dx}{\int_0^L \rho dx} \quad \int_0^L \rho dx = 1$$



$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$\mu = \int_{-\infty}^{\infty} x f(x) dx$$

$$\mu' = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

Moments and Moment Generating Functions

Definition

Let X be any random variable with pdf $f_X(x)$. For any positive integer r:

- ***The r-th moment of X about the origin**, μ_r, is given by*

$$\mu_r = E[X^r] = \int_{-\infty}^{\infty} w^r \cdot f_X(w)dw,$$

provided $\int_{-\infty}^{\infty} |w|^r \cdot f_W(w)dw < \infty$.
(When $r = 1$, the subscript is usually omitted, i.e. $\mu_1 = \mu$.)

DEFINITION

Let X be any random variable with pdf $f_X(x)$. For any positive integer r:

- ***The r-th moment of X about the origin**, μ_r, is given by*

$$\mu_r = E[X^r] = \int_{-\infty}^{\infty} w^r \cdot f_X(w)dw,$$

  provided $\int_{-\infty}^{\infty} |w|^r \cdot f_W(w)dw < \infty$.
  (When $r = 1$, the subscript is usually omitted, i.e. $\mu_1 = \mu$.)

- ***The r-th moment of X about the mean,** μ'_r, is given by*

$$\mu'_r = E\left((X-\mu)^r\right) = \int_{-\infty}^{\infty} (w-\mu)^r \cdot f_X(w)dw,$$

  provided fitness conditions of the part 1 hold.
  (When r=2, $\mu'_2 = \sigma^2$ is called the `variance`, and σ is called the `standard deviation`.)

MOMENTS AND MOMENT GENERATING FUNCTIONS

DEFINITION (CONTINUED...)

- ***The** $r-$**th standardized moment**, $\tilde{\mu}_r$, is a moment that is normalized, typically by the standard deviation raised to the power of r, σ^r.*

$$\tilde{\mu}_r = \frac{E\left((W-\mu)^r\right)}{\sigma^r}.$$

Moments and Moment Generating Functions

$E[X]$ - the *mean* - the first moment about origin:

$$\mu = E[X] = \int_{-\infty}^{\infty} w \cdot f_X(w)dw$$

$$\mu = E[X] = \sum_{\text{all } x} x \cdot p_X(x)$$

$E[(X-\mu)^2]$ - the *variance* - the second moment about the mean;

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (w-\mu)^2 \cdot f_X(w)dw$$

$$\sigma^2 = E[(X-\mu)^2] = \sum_{\text{all } x} (x-\mu)^2 \cdot p_X(x)$$

$E[(X-\mu)^3]/(E[(X-\mu)^2])^{3/2}$ - the *skewness* - the third standardized moment;
$E[(X-\mu)^4]/(E[(X-\mu)^2])^2$ - the *kurtosis* - the forth standardized moment.

MOMENTS AND MOMENT GENERATING FUNCTIONS

EXAMPLE

To find out the prevalence of smallpox vaccine use, a researcher inquired into the number of times a randomly selected 200 people aged 16 and over in an African village had been vaccinated. He obtained the following figures:

N	0	1	2	3	4	5
proportion	17/200	30/100	58/200	50/200	38/200	7/200

Assume that these proportions continue to hold exhaustively for the population of that village.

Questions:

a) What is the expected number of times those people in the village had been vaccinated?

b) What is the standard deviation?

EXAMPLE (SOLUTION)

The random variable X represent the number of times a person has been vaccinated, then X has distribution p(x) as in the table.

x	0	1	2	3	4	5
p(x)	17/200	30/200	58/200	50/200	38/200	7/200

Then

$$E[X] = \sum xp(x) = 0 \cdot \frac{17}{200} + 1 \cdot \frac{30}{200} + 2 \cdot \frac{58}{200} \dots + 5 \cdot \frac{7}{200} = \frac{483}{200} = 2.415$$

The variance

$$\sigma^2 = Var[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2 =$$

$$= 0 \cdot \frac{17}{200} + 1^2 \cdot \frac{30}{200} + 2^2 \cdot \frac{58}{200} + \dots + 5^2 \cdot \frac{7}{200} - (2.415)^2 = 1.643 \text{ (3 d.p.)}$$

Example

Let X be a random variable with pdf:

$$f(x) = \begin{cases} \frac{3}{64}y^2(4-y) & 0 \leq x \leq 4 \\ 0, & \text{otherwise} \end{cases} .$$

Question:

Find the expected value and variance of X.

In Problem Sheet 1

MOMENTS AND MOMENT GENERATING FUNCTIONS

EXAMPLE

Find the expectation and the variance for a random variable X:

- with Poisson distribution, $Poi(\lambda)$:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \text{ for } k = 0, 1, 2, \ldots$$

$$E(k) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \cdot \lambda \overset{t=k-1}{=} e^{-\lambda} \lambda \cdot \underbrace{\sum_{t=0}^{\infty} \frac{\lambda^t}{(t-1)!}}_{= e^{\lambda}}$$

$$\Rightarrow \underline{\underline{E(k) = e^{-\lambda} \cdot e^{\lambda} \cdot \lambda = \lambda}}$$

- with normal distribution, $N(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for } -\infty < x < \infty$$

was discussed in Probability module

Definition

For a random variable X, suppose that there is a positive number h such that for $-h < t < h$ the mathematical expectation $E(e^{tX})$ exists. The ***moment-generating function*** *(mgf) of the random variable X is defined by*

$$M_X = E(e^{tX}) = \sum_{-\infty}^{\infty} e^{tk} p_X(k), \quad \textit{if } X \textit{ is discrete}$$

$$M_X = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx, \quad \textit{if } X \textit{ is continuous.}$$

MOMENT-GENERATING FUNCTIONS

$$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + ... + \frac{(tX)^n}{n!} + ...$$

The moment generating function

$$M_X = E(e^{tX}) = 1 + tE(X) + \frac{(t)^2}{2!}E(X^2) + ... + \frac{(t)^n}{n!}E(X^n) + ...$$

Moment-Generating functions

$$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + ... + \frac{(tX)^n}{n!} + ...$$

The moment generating function

$$M_X = E(e^{tX}) = 1 + tE(X) + \frac{(t)^2}{2!}E(X^2) + ... + \frac{(t)^n}{n!}E(X^n) + ...$$

$$\frac{dM_X}{dt} = M'_X(t) = E(X) + tE(X^2) + ... + \frac{(t)^{n-1}}{(n-1)!}E(X^n) + ...$$

MOMENT-GENERATING FUNCTIONS

$$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + ... + \frac{(tX)^n}{n!} + ...$$

The moment generating function

$$M_X = E(e^{tX}) = 1 + tE(X) + \frac{(t)^2}{2!}E(X^2) + ... + \frac{(t)^n}{n!}E(X^n) + ...$$

$$\frac{dM_X}{dt} = M_X'(t) = E(X) + tE(X^2) + ... + \frac{(t)^{n-1}}{(n-1)!}E(X^n) + ...$$

Evaluating for $t = 0$

$$M_X'(0) = E(X), \qquad M_X''(0) = E(X^2), \qquad ... M_X^{(n)}(0) = E(X^n), \; n \geq 1$$

MOMENT-GENERATING FUNCTIONS

THEOREM

If $M_X(t)$ exists, then for any positive integer r,

$$\left.\frac{d^r M_X}{dt^r}\right|_{t=0} = M_X^{(r)}(0) = \mu_r$$

Moment-Generating functions

Theorem

If $M_X(t)$ exists, then for any positive integer r,

$$\left.\frac{d^r M_X}{dt^r}\right|_{t=0} = M_X^{(r)}(0) = \mu_r$$

Properties of the moment-generating function:

MOMENT-GENERATING FUNCTIONS

THEOREM

If $M_X(t)$ exists, then for any positive integer r,

$$\left.\frac{d^r M_X}{dt^r}\right|_{t=0} = M_X^{(r)}(0) = \mu_r$$

Properties of the moment-generating function:

1. The moment-generating function of X is unique in the sense that, if two random variables X and Y have the same mgf ($M_X(t) = M_Y(t)$, for t in an interval containing 0), then X and Y have the same distribution.

MOMENT-GENERATING FUNCTIONS

THEOREM

If $M_X(t)$ exists, then for any positive integer r,

$$\left.\frac{d^r M_X}{dt^r}\right|_{t=0} = M_X^{(r)}(0) = \mu_r$$

Properties of the moment-generating function:

1. The moment-generating function of X is unique in the sense that, if two random variables X and Y have the same mgf ($M_X(t) = M_Y(t)$, for t in an interval containing 0), then X and Y have the same distribution.
2. If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$. That is, the mgf of the sum of two independent random variables is the product of the mgfs of the individual random variables. The result can be extended to n random variables.

Moment-generating functions

Theorem

If $M_X(t)$ exists, then for any positive integer r,

$$\left.\frac{d^r M_X}{dt^r}\right|_{t=0} = M_X^{(r)}(0) = \mu_r$$

Properties of the moment-generating function:

1. The moment-generating function of X is unique in the sense that, if two random variables X and Y have the same mgf ($M_X(t) = M_Y(t)$, for t in an interval containing 0), then X and Y have the same distribution.
2. If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$. That is, the mgf of the sum of two independent random variables is the product of the mgfs of the individual random variables. The result can be extended to n random variables.
3. Let $Y = aX + b$. Then $M_Y(t) = e^{bt}M_X(at)$.

QUESTIONS TO TAKE HOME

1. Difference between Descriptive and Inferential Statistics;

Questions to take home

1. Difference between Descriptive and Inferential Statistics;
2. Difference between Population and Sampling, between Parameter and Statistic;

QUESTIONS TO TAKE HOME

1. Difference between Descriptive and Inferential Statistics;
2. Difference between Population and Sampling, between Parameter and Statistic;
3. Different ways of sampling(simple random sampling, cluster sampling, and stratified sampling);

Questions to take home

1. Difference between Descriptive and Inferential Statistics;
2. Difference between Population and Sampling, between Parameter and Statistic;
3. Different ways of sampling(simple random sampling, cluster sampling, and stratified sampling);
4. Find some examples of descriptive and inferential statistics.