

MATH1905 Statistics

Lecture 13

Lecturer: Marc Raimondo
Carlaw 817

Lectures: Mon.11am (Chem LT1), Tue.8am (Chem LT1)

Student consultations: Tuesday 2-3pm

General information, tutes, solutions etc...

<http://www.maths.usyd.edu.au/u/UG/JM/MATH1905/>

or

First Year Office (FYO), Carlaw 520.

Lecture 13: Independent rv's.

Discrete rv's X and Y are defined to be *independent* if the events $\{X = x\}$ and $\{Y = y\}$ are independent for *all possible values* of x and y .

Example. IN the following experiment, are X and Y independent?

Let X = number of successes, Y = number of failures in

(a) n Bernoulli trials,

$$P(X = 0) = (1 - p)^n, \quad P(Y = 0) = p^n,$$

$$\text{But } P(X = 0, Y = 0) = 0 \neq P(X = 0) \times P(Y = 0)$$

so X and Y are NOT INDEPENDENT.

(b) N Bernoulli trials, where $N \sim \mathcal{P}(\lambda)$.

$$P(X = 0 | N = n) = (1 - p)^n$$

Applying the total probability rule

$$P(X = 0) = \sum_{n \geq 0} P(X = 0 | N = n) \times P(N = n)$$

$$P(X = 0) = \sum_{n \geq 0} (1 - p)^n \frac{\lambda^n}{n!} e^{-\lambda}$$

$$P(X = 0) = e^{(1-p)\lambda} e^{-\lambda} = e^{-p\lambda}.$$

Similarly $P(Y = 0) = e^{-p\lambda}$.

but $P(X = 0, Y = 0) = 0$. X, Y NOT INDEPENDENT

Equivalent definition: the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for all x and y . In the continuous case, it follows that if X and Y are independent, with pdf's $f(x)$ and $g(y)$, then

$$P(x < X < x + dx, y < Y < y + dy) \approx f(x) g(y) dx dy$$

Extension to more than 2

rv's:
$$P(x_1 < X_1 < x_1 + dx_1, x_2 < X_2 < x_2 + dx_2, \dots, x_n < X_n < x_n + dx_n) \approx f(x_1) f(x_2) \dots f(x_n) dx_1 \dots dx_n$$

with dist. function:

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1) \dots P(X_n \leq x_n)$$

Sums of independent random variables.

Suppose that X and Y are integer-valued rv's. The means of the sum and product of X and Y are defined to be

$$\mathbf{E}(X + Y) = \sum_{i,j} \sum (i + j) \mathbf{P}(X = i, Y = j)$$

$$\mathbf{E}(XY) = \sum_{i,j} \sum i j \mathbf{P}(X = i, Y = j)$$

These results are particular cases of the general

formula: $\mathbf{E}(g(X, Y)) = \sum \sum_{i,j} g(i, j) \mathbf{P}(X = i, Y = j)$

Expectation of Sums of rv's

For any rv's X and Y , independent or not, the mean of the sum is the sum of the means, that is,

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y).$$

$$\begin{aligned} \textit{Proof.} \mathbf{E}(X + Y) &= \sum \sum_{i,j} (i + j) \mathbf{P}(X = i, Y = j) \\ &= \sum_i \sum_j (i + j) \mathbf{P}(X = i | Y = j) \mathbf{P}(Y = j) \\ &= \sum_i (\sum_j i \mathbf{P}(X = i | Y = j) \mathbf{P}(Y = j)) + \\ &\quad + \sum_j (\sum_i j \mathbf{P}(Y = j | X = i) \mathbf{P}(X = i)) \\ &= \sum_i i \mathbf{P}(X = i) + \sum_j j \mathbf{P}(Y = j) = \mathbf{E}X + \mathbf{E}Y \end{aligned}$$

Variance of Sums of rv's

For independent rv's X and Y ,

$$\mathit{var}(X + Y) = \mathit{var}(X) + \mathit{var}(Y).$$

Proof. We suppose that $EX = EY = 0$.

$$E(XY) = \sum_{i,j} ijP(X = i, Y = j)$$

$$\text{using independence} = \sum_{i,j} ijP(X = i)P(Y = j)$$

$$E(XY) = \sum_j \left(\sum_i iP(X = i) \right) j P(Y = j)$$

$$= \sum_j (E(X)j P(Y = j)) = E(X) \sum_j (j P(Y = j)) = E(X)E(Y) = 0$$

$$\text{Now } \mathit{var}(X + Y) = E(X + Y)^2 = E(X^2 + Y^2 + 2XY) = EX^2 + EY^2 + E(XY)$$

$$\text{But } E(XY) = E(X)E(Y) = 0 \text{ so}$$

$$\mathit{var}(X + Y) = EX^2 + EY^2 = \mathit{var}(X) + \mathit{var}(Y)$$

We have used the fact that

$$E(X) = 0 \rightarrow \mathit{var}(X) = E(X - E(X))^2 = EX^2$$

Examples.

- If X and Y are independent rv's then

$$\mathit{var}(aX + bY) = a^2\mathit{var}(X) + b^2\mathit{var}(Y).$$

In particular, $\mathit{var}(X - Y) = \mathit{var}(X) + \mathit{var}(Y)$.

Solution. $E(aX + bY) = E(a^2X^2 + b^2Y^2 + 2abXY) =$
 $a^2\mathit{var}(X) + b^2\mathit{var}(Y) + 2abE(XY)$

$$= a^2\mathit{var}(X) + b^2\mathit{var}(Y) + 0$$

Taking $a = 1$ $b = -1$

$$\mathit{var}(X - Y) = \mathit{var}(X) + \mathit{var}(Y).$$

- If X_1, \dots, X_n are independent, all with mean μ and variance σ^2 , then

$$\mathbf{E} \left(\sum_{i=1}^n X_i \right) = n\mu \text{ and } \mathit{var} \left(\sum_{i=1}^n X_i \right) = n\sigma^2.$$

The proof is left as an exercise.

Sums of independent normal variables

In general it's hard to find the *distribution* of a sum of independent rv's. However, the sum of independent *normal* variables gives another *normal* rv:

if X_1 and X_2 are independent, $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

This generalises to *weighted* sums of independent rv's: let a_1, \dots, a_n be constants and independent rv's $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, 2, \dots, n$ then

$$\sum_{i=1}^n a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right).$$

Sums of iid rv's

If X_1, \dots, X_n are independent rv's with the same distribution, we say they are *iid* (independent and identically distributed).

In the special case where X_1, \dots, X_n are iid with a normal $\mathcal{N}(\mu, \sigma^2)$ distribution, it follows that

- $\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$
- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Examples. 1. Steel rods, made with diameter R normally distributed about 4.90cm, with standard deviation 0.03cm, are to fit into sockets, made with diameter S normally distributed about 5.00cm with standard deviation 0.04cm. For a satisfactory fit the socket diameter should exceed the rod diameter, but by not more than 0.20cm. If a rod and a socket are taken at random, what is the probability that the fit is unsatisfactory?

Solution. $R \sim \mathcal{N}(4.9, 0.03^2)$, $S \sim \mathcal{N}(5, 0.04^2)$

$S - R \sim \mathcal{N}(5 - 4.9, 0.04^2 + 0.03^2)$

$S - R \sim \mathcal{N}(0.1, 0.05^2)$

$P(S - R > 0.2) = P(\mathcal{N}(0.1, 0.05^2) > 0.2) =$

$> 1 - \text{pnorm}(0.2, 0.1, 0.05)$

[1] 0.02275013

Examples. 2. The examination scores in a certain university course are approximately normally distributed with mean 56 and standard deviation 11.

In a class of 49 students, what is the probability that the average mark is less than 50? What is the probability that the average mark lies between 50 and 60?

Solution. $X_i \sim \mathcal{N}(56, 11^2)$, $Z \sim \mathcal{N}(0, 1)$

$$\bar{X} \sim \mathcal{N}\left(56, \frac{11^2}{49}\right)$$

$$P(50 < \bar{X} < 60) = P\left(\frac{50-56}{\frac{11}{7}} < Z < \frac{60-56}{\frac{11}{7}}\right)$$

$$P(50 < \bar{X} < 60) = P(-3.8 < Z < 3.8) \approx 1$$

$$P(\bar{X} < 50) = P(Z < -3.8) \approx 0$$

Review problems. P&Q pp. 79–80: 15, 16, 17.

R commands: A loop to simulate 100 observations of a $B(75, 0.5)$ distribution.

```
vector.binomial=numeric(0)      # a place to store the result
for (i in 1:100){                # for the loop
vector.binomial[i]=rbinom(1,75,0.5)  # observe 1 B(75,0.5)
}
```

A loop to simulate 100 vectors of size 200 of a uniform distribution

```
matrix.uniform=matrix(0,100,200)  # matrix with 100 rows 200 cols
for (i in 1:100){                 # for the loop
matrix.uniform[i,]=runif(200)     # observe 200 U(0,1)
}
```

R command for finding means by rows

```
means.uniform=apply(matrix.uniform,1,mean)
```

Lecture 14: Sampling distributions.

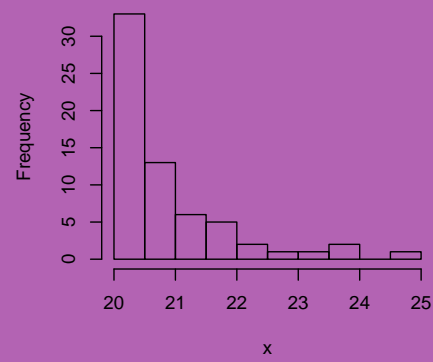
In terms of random variables, a random sample of size n is a set of n independently obtained values of a random variable X , which is sometimes called the 'parent variable'. The observed sample, x_1, x_2, \dots, x_n , will be regarded as one of many possible samples.

A second sample of size n , taken under the same conditions, would result in another set of n values of X : x'_1, x'_2, \dots, x'_n and so on $x''_1, x''_2, \dots, x''_n$. The values x_1, x_2, \dots, x_n are independent observations of the parent variable X , but so are x_i, x'_i, x''_i, \dots , and in the hypothetical population of possible samples, the i th sample value is represented by a random variable, X_i . The n rv's X_1, X_2, \dots, X_n are iid, with the same distribution as the parent variable X .

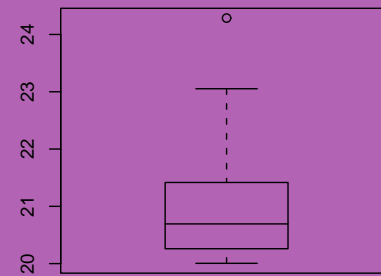
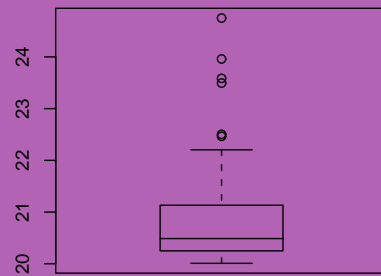
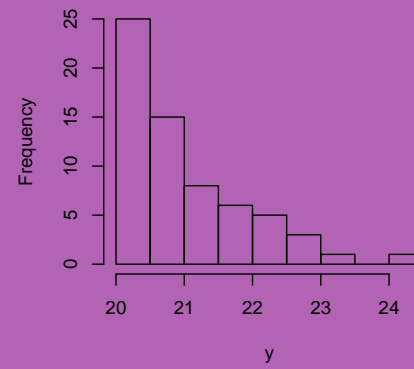
Example

Histograms and boxplots of two random samples (x_1, \dots, x_{64}) and (x'_1, \dots, x'_{64}) , each consisting of $n = 64$ observations of a random variable with mean 21 and variance 1:

Histogram of x



Histogram of y



Statistics

Suppose we have a sample (x_1, \dots, x_n) from a population with parent variable X . Any function of the sample values x_1, x_2, \dots, x_n is called a *statistic*.

Any statistic, calculated from the sample values, can be thought of as the observed value of a rv.

Consider the statistic \bar{x} , the sample mean. To \bar{x} there corresponds the rv $\bar{X} = \frac{1}{n} (X_1 + \dots + X_n)$, whose probability distribution (the 'sampling distribution') obviously depends on the distribution of the parent variable X .

If X has mean μ and variance σ^2 , then

$$\mathbf{E}(\bar{X}) = \frac{1}{n} \mathbf{E}(X_1 + \cdots + X_n) = \mu \quad \text{and}$$

$$\text{var}(\bar{X}) = \text{var}\left(\frac{X_1}{n} + \cdots + \frac{X_n}{n}\right) = \frac{\sigma^2}{n^2} + \cdots + \frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

The standard deviation of a statistic is called its *standard error*. Thus the sample mean \bar{X} is distributed about its mean $\mathbf{E}(\bar{X}) = \mu$, with standard error σ/\sqrt{n} .

The central limit theorem

In the case where the parent variable is *normal* (i.e. the sample is from a normal population) it follows from the additive property of independent normal variables that \bar{X} is also a normal variable:

$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. The Central Limit Theorem says that even if the parent variable is *not normal*, \bar{X} is *approximately normally distributed*, so long as n is large. In the following sense:

If X_1, \dots, X_n are iid rv's with common mean μ and common variance $\sigma^2 > 0$, then the distribution function of the standardized variable

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}}$$

tends to the standard normal distribution function as $n \rightarrow \infty$.

Equivalent versions of the CLT:

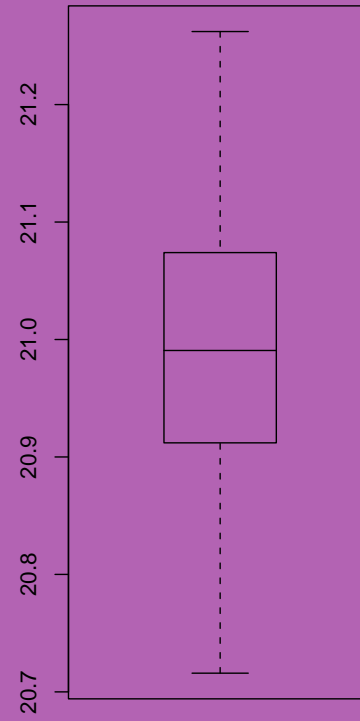
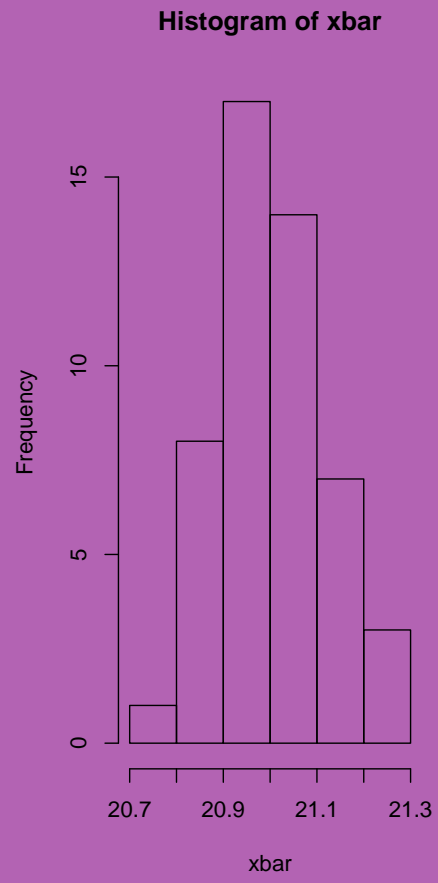
$\sum_{i=1}^n X_i$ is approximately normal $\mathcal{N}(n\mu, n\sigma^2)$

\bar{X} is approximately normal $\mathcal{N}(\mu, \frac{\sigma^2}{n})$

The approximation is good for *any distribution* if $n > 25$ and even for smaller values of n if the distribution of the X_i 's is fairly symmetric and doesn't have long tails.

Example: we generated 50 data sets of size $n = 64$ from a (**exponential**) distribution with mean $\mu = 21$ and variance $\sigma^2 = 1$. (First two shown on Slide 2.) They are clearly not from a normal distribution.

Histogram and boxplot of 50 sample means



The 50 values have a sample mean of **20.98** and a sample standard deviation of **0.12**; the corresponding rv \bar{X} has mean (expected value) **$\mu = 21$** and standard error **$\sigma/\sqrt{64} = 0.125$** , because the X_i 's had mean **$\mu = 21$** and variance **$\sigma^2 = 1$** .

The histogram shows that the distribution of \bar{X} appears normal...even though the parent distribution was not normal.

Normal approximation to binomial

From the definition of $X \sim \mathcal{B}(n, p)$ as the number of successes in n independent trials with success probability p , it follows that we can write $X = \sum_{i=1}^n X_i$ where X_1, \dots, X_n are iid with $\mathbf{P}(X_i = 1) = p, \mathbf{P}(X_i = 0) = 1 - p$.

Because of this representation of X as a sum of iid rv's, the **central limit theorem** implies that for large n , X will be approximately normal $\mathcal{N}(np, np(1 - p))$. The approximation being better for p not too far from $1/2$ to ensure approximate symmetry. (As a guide, the approximation may not be very good unless $n > 25, np > 5$ and $n(1 - p) > 5$.)

Continuity correction

We can exploit the fact that X takes only integer values to improve the approximation by using a *continuity correction*. Let $Y \sim \mathcal{N}(np, np(1-p))$ be the approximating normal variable. Then the CLT says we can use the approximation $P(X \leq j) \approx P(Y \leq j)$. But the following diagram shows that

$P(X \leq j) \approx P(Y \leq j + \frac{1}{2}) = \Phi\left(\frac{j + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$ may be an improvement.

Continuity correction...Similarly

$$\mathbf{P}(X \geq j) \approx \mathbf{P}(Y \geq j - 1/2) = \mathbf{1} - \Phi \left(\frac{j - 1/2 - np}{\sqrt{np(1-p)}} \right)$$

$$\begin{aligned} \mathbf{P}(i \leq X \leq j) &\approx \mathbf{P}(i - 1/2 \leq Y \leq j + 1/2) \\ &= \Phi \left(\frac{j + 1/2 - np}{\sqrt{np(1-p)}} \right) - \Phi \left(\frac{i - 1/2 - np}{\sqrt{np(1-p)}} \right) \end{aligned}$$

$$\mathbf{P}(X < j) \approx \mathbf{P}(Y \leq j - 1/2) = \Phi \left(\frac{j - 1/2 - np}{\sqrt{np(1-p)}} \right)$$

Examples.

1. Find the probability that among 10,000 random digits the digit 7 appears no more than 968 times.

Solution. For 1 digit $\Omega = \{0, 1, 2, \dots, 9\}$ and $P(7) = \frac{1}{10}$

Let $X = \#7's$ in a sample of 10,000 digits: $X \sim B(10000, \frac{1}{10})$

The approximating normal is $Y \sim \mathcal{N}(1000, 900)$

with continuity correction:

$$P(X \leq 968) \approx P(Y \leq 968.5)$$

$$P(X \leq 968) \approx P(Z \leq \frac{968.5 - 1000}{\sqrt{900}})$$

$$P(X \leq 968) \approx P(Z \leq 1.05) = 0.8577(4dp)$$

Examples.

2. Find a number k such that the probability is about $1/2$ that the number of heads obtained in 1000 tossings of a coin will be between 490 and k .

Solution. $P(490 \leq X \leq k) = 0.5$

$X \sim B(1000, \frac{1}{2})$. The approximating normal is $Y \sim \mathcal{N}(500, 250)$ without continuity correction:

$$P\left(\frac{490-500}{\sqrt{250}} \leq Z \leq \frac{k-500}{\sqrt{250}}\right) = 0.5$$

$$\Phi\left(\frac{k-500}{\sqrt{250}}\right) - \Phi\left(\frac{-10}{\sqrt{250}}\right) = 0.5$$

$$\Phi\left(\frac{k-500}{\sqrt{250}}\right) = 0.5 + \Phi\left(\frac{-10}{\sqrt{250}}\right)$$

$$\Phi\left(\frac{k-500}{\sqrt{250}}\right) = 0.7635(4dp)$$

In R `qnorm(0.7635)` gives **0.7176** so we find that

$$k = 0.7176 \times \sqrt{250} + 500 \approx 511$$

Examples.

3. A sample is taken in order to find the fraction f of females in a population. Find a sample size such that the probability of a sampling error less than 0.005 will be 0.99 or greater.

Solution. $P(F) = f$, $P(M) = 1 - f$.

Sampling (X_i) iid, $X_i = 1$ if F and $X_i = 0$ if M .

$$T = \sum_{i=1}^n X_i \sim B(nf, nf(1 - f)) \approx \mathcal{N}(nf, nf(1 - f))$$

Estimating f by $\hat{f} = \bar{X} \approx \mathcal{N}(f, f(1 - f)/n)$

We want to find n such that

$$P(|\bar{X} - f| \leq 0.005) \geq 0.99$$

$$\bar{X} - f \sim \mathcal{N}(0, f(1-f)/n)$$

$$P(|Z| \leq \frac{0.005}{\sqrt{\frac{f(1-f)}{n}}}) \geq 0.99$$

$$P(|Z| \geq \frac{0.005}{\sqrt{\frac{f(1-f)}{n}}}) \leq 0.001$$

$$P(|Z| \geq \frac{0.005}{\sqrt{\frac{f(1-f)}{n}}}) \leq P(|Z| \geq \frac{0.005}{\sqrt{\frac{1}{4n}}})$$

$$P(|Z| \geq 0.005 \times 2 \times \sqrt{n}) \leq 0.001$$

$$\text{by symmetry } 2 \times (1 - \Phi(0.001\sqrt{n})) \leq 0.001$$

$$\Phi(0.001\sqrt{n}) \geq 0.9995$$

In R `qnorm(0.9995)` gives 3.291 so $n \approx 108306.8$

Review problems P&Q pp. 80–81: 18, 19.

R commands to illustrate CLT

A loop to simulate 1000 observations of a $B(1000, 0.5)$ distribution.

```
vector.clt=numeric(0)      # a place to store the result
for (i in 1:1000){         # for the loop
  bin.obs=rbinom(1,1000,0.5) # observe 1 B(1000,0.5)
  vector.clt[i]=(bin.obs-500)/sqrt(250) # CLT-standardisation
}
```

Check the distribution of `vector.clt`: `hist(vector.clt)`
`qqnorm(vector.clt)` `plot(density(vector.clt))`

Do a similar experiment where `vector.clt` is a vector of 1000 sample means obtained from 1000 samples of size 1000 of a $U(0, 1)$ distribution.