UNSW Mathematics Society presents

MATH2089/2099/2859

CVEN2002 Revision Seminar

Statistics

T2, 2020

Presented by Gerald Huang

Table of Contents I

Table of Contents II

Part I: Random variables

Random variable

Definition I: Random variable

A random variable is a real-valued function defined over the sample space $X : S \to \mathbb{R}$ and $\omega \to X(\omega)$.

Cumulative distribution function (CDF)

Definition: Cumulative distribution function

A **cumulative distribution function** of a random variable X is defined, for any real number x, as

$$F(x) = \mathbb{P}(X \leq x).$$

Properties.

- For any real numbers $a \leq b$, we have

$$\mathbb{P}(a < X \leq b) = F(b) - F(a).$$

- It is **nondecreasing**. That is, if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
- $\lim_{x \to +\infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$.

Discrete Random Variables

Definition: Discrete Random Variables

A random variable is said to be discrete if it can only assume a finite (or at most countably infinite) number of values.

- Essentially we can count each event!

Characterising a discrete random variable

Discrete random variables can be characterised by their **probability mass function** (pmf), defined by

$$p(x) = \mathbb{P}(X = x).$$

- The sum of ALL elements x in the event A is 1. That is,

$$\sum_{x \in A} p(x) = 1.$$

Continuous Random Variables

Definition: Continuous Random Variables

A random variable is said to be continuous if it is defined over an **uncountable** set of real numbers, usually an intervals.

Characterising a continuous random variable

Continuous random variables can be characterised by their **probability density function** (pdf), defined by $f(x)$.

- The integral over ALL elements x in the event space A is 1. That is,

$$\int_A f(x)\, dx = 1.$$

Example

To determine whether $f(x) = e^{-x}$ for $x > 0$ is a density function, check whether

$$\int_0^{\infty} e^{-x}\,dx = 1.$$

Expectation of random variables

Expectation of a discrete random variable

The expectation (or mean) of a discrete random variable, denoted $\mathbb{E}(X)$ or μ, is defined by

$$\mu = \mathbb{E}(X) = \sum_{x \in A} xp(x).$$

Expectation of a continuous random variable

The expectation (or mean) of a continuous random variable, denoted $\mathbb{E}(X)$ or μ, is defined by

$$\mu = \mathbb{E}(X) = \int_A xf(x)\,dx.$$

Expectation of random variables (18S2)

Example: (2018 Semester 2, Q3a)

Let X follow the Bernoulli distribution:

$$p(x) = \begin{cases} 1-\pi, & \text{if } x = 0 \\ \pi, & \text{if } x = 1 \end{cases}$$

where $0 < \pi < 1$.
Show that $\mathbb{E}(X) = \pi$.

Since this is a **discrete** random variable, then the expected value is simply

$$\mathbb{E}(X) = \sum_{x \in X} xp(x) \;= 0 \times (1-\pi) + 1 \times \pi \;= \pi.$$

Properties of the expectation function

- **Linearity**: For any two constants a and b, we have

$$\mathbb{E}(aX + b) = a \cdot \mathbb{E}(X) + b.$$

- **Degenerate**: A random variable X is said to be degenerate if

$$\mathbb{E}(b) = b.$$

Example

If $\mathbb{E}(X) = 2$, then

$$\mathbb{E}(3X + 4) = 3 \times \mathbb{E}(X) + 4 = 3 \times 2 + 4 = 10.$$

Example

If $\mathbb{E}(3X + 4) = 10$, then $3\mathbb{E}(X) + 4 = 10 \implies \mathbb{E}(X) = 2$.

Variance of a random variable

Variance of a random variable

The **variance** of a random variable, denoted by $\mathrm{Var}(X)$ or σ^2, is defined by

$$\mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

Properties of the variance function

- For any random variable, $\mathrm{Var}(X) \geq 0$.
- For any two constants a and b, $\mathrm{Var}(aX+b) = a^2 \cdot \mathrm{Var}(X)$.
- For any constant b, $\mathrm{Var}(b) = 0$.

Computing the variance

Variance of a discrete random variable

The variance of a discrete random variable is defined by

$$\mathrm{Var}(X) = \sum_{x \in A} (x-\mu)^2 p(x) = \underbrace{\left(\sum_{x \in A} x^2 p(x)\right)}_{\mathbb{E}(X^2)} - \underbrace{\left(\sum_{x \in A} x p(x)\right)^2}_{\mathbb{E}(X)^2}$$

Variance of a continuous random variable

The variance of a continuous random variable is defined by

$$\mathrm{Var}(X) = \int_A (x-\mu)^2 f(x)\, dx = \underbrace{\left(\int_A x^2 f(x)\, dx\right)}_{\mathbb{E}(X^2)} - \underbrace{\left(\int_A x f(x)\, dx\right)^2}_{\mathbb{E}(X)^2}$$

Example

If $f(x) = e^{-x}$ for $x > 0$, then the variance can be found by computing the integral

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \int_0^\infty x^2 e^{-x}\, dx - \left(\int_0^\infty x e^{-x}\, dx\right)^2$$

Standard deviation

- The **standard deviation** is simply the square root of the variance. That is,
$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$
- Since $\mathrm{Var}(X) \geq 0$, then the standard deviation function will always be defined!

Jointly distributed random variables

- We will now turn towards the two-dimensional case and discuss properties of distributions of *two* random variables!

Joint cumulative distribution function

Definition: Joint cumulative distribution function (discrete)

The joint cumulative distribution function of discrete random variables X and Y is given by

$$F_{XY}(x, y) = \mathbb{P}(X \leq x, Y \leq y), \qquad \text{for all } (x, y) \in \mathbb{R} \times \mathbb{R}.$$

Definition: Joint cumulative distribution function (continuous)

X and Y are said to be jointly continuous if, for any sets A and B of real numbers, there is a function (the joint probability density of X and Y) $f_{XY}(x, y)$

$$\mathbb{P}(X \in A, Y \in B) = \int_A \int_B f_{XY}(x, y)\, dy\, dx.$$

Joint distribution functions and marginal functions

Discrete

Joint distribution

$$p_{XY}(x, y) = \mathbb{P}(X = x, Y = y).$$

Marginal probabilities

$$p_X(x) = \sum_{y \in S_Y} p_{XY}(x, y).$$

$$p_Y(y) = \sum_{x \in S_X} p_{XY}(x, y).$$

Continuous

Joint distribution

Denoted as $f_{XY}(x, y)$.

Marginal densities

$$f_X(x) = \int_{S_Y} f_{XY}(x, y)\, dy.$$

$$f_Y(y) = \int_{S_X} f_{XY}(x, y)\, dx.$$

Expectation of a function of two random variables

For any function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the expectation of $g(X, Y)$ is given by

$$\mathbb{E}(g(X, Y)) =$$

Discrete random variables

$$\sum_{x \in S_X} \sum_{y \in S_Y} g(x, y) p_{XY}(x, y)$$

Continuous random variables

$$\int_{S_X} \int_{S_Y} g(x, y) f_{XY}(x, y)\, dy\, dx$$

Linearity property of the expectation function still holds!

$$\mathbb{E}(aX + bY) = a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y).$$

Example: Table of marginal probabilities

	0	1	2	3
-1	1/8	1/8	1/8	1/8
1	1/8	1/4	1/2	5/8
2	1/8	3/8	3/4	7/8
3	1/8	1/2	7/8	1

Assume that X is across the top and Y is on the side. Find $\mathbb{P}(X \leq 1, Y \leq 1)$.

$$\begin{aligned}
&\mathbb{P}(X \leq 1, Y \leq 1) \\
&= \mathbb{P}(X = 0, Y = -1) + \mathbb{P}(X = 0, Y = 1) \\
&+ \mathbb{P}(X = 1, Y = -1) + \mathbb{P}(X = 1, Y = 1) \\
&= 1/8 + 1/8 + 1/8 + 1/4 = 5/8.
\end{aligned}$$

Independent random variables

Definition: Independence of random variables

Random variables X and Y are said to be **independent** if, for all $(x, y) \in \mathbb{R} \times \mathbb{R}$,

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \times \mathbb{P}(Y \leq y).$$

Discrete case

$$p_{XY}(x, y) = p_X(x) \times p_Y(y).$$

Continuous case

$$f_{XY}(x, y) = f_X(x) \times f_Y(y).$$

Property of independent random variables

If X and Y are **independent**, then for any functions h and g,

$$\mathbb{E}(h(X)g(Y)) = \mathbb{E}(h(X)) \times \mathbb{E}(g(Y)).$$

Example: (MATH2089, 2009S1 Q5c)

Suppose that X and Y are independent standard normal variables. What is the distribution of $X + Y$?

Since X and Y are independently and normally distributed , then their sum is also normally distributed with

$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2, \sigma_Y^2) = \mathcal{N}(0, 2).$$

Covariance of two random variables

Definition: Covariance of two random variables

The **covariance** of two random variables X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right].$$

Properties of covariance

- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- **Symmetric**: For any two variables X and Y, $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
- **IMPORTANT**: $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.
- $\mathrm{Cov}(aX + b, cY + d) = ac\,\mathrm{Cov}(X, Y)$
- **Bilinearity**: $\mathrm{Cov}(X_1 + X_2, Y_1 + Y_2) = \mathrm{Cov}(X_1, Y_1) + \mathrm{Cov}(X_1, Y_2) + \mathrm{Cov}(X_2, Y_1) + \mathrm{Cov}(X_2, Y_2)$.

Covariance and independence

- If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. But if $\mathrm{Cov}(X, Y) = 0$, then X and Y may or may not be independent!

Remark

X and Y independent $\implies \mathrm{Cov}(X, Y) = 0$.
$\mathrm{Cov}(X, Y) = 0 \;\not\Longrightarrow\; X$ and Y independent.

Variance of a sum of random variables

Variance of a sum of two random variables

For any two random variables X and Y,

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y).$$

- If X and Y are independent, then

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y).$$

Correlation coefficient

Definition: Correlation

The correlation coefficient denoted by ρ is defined as

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

- We are computing the covariance between the **standardised** versions of X and Y.

Properties of correlation

- ρ does not have a unit.
- $-1 \leq \rho \leq 1$.
- Positive ρ means positive linear relationship between X and Y and vice versa for negative!
- The closer $|\rho|$ is to 1, the stronger the relationship!

Part II: Sampling distributions and Central Limit Theorem

Independent and identically distributed random variables

A sequence of random variables $X_1, X_2, \ldots, X_N$ are said to be *i.i.d* if

1. all X_i's are independent.
2. all X_i's share the same probability distribution (identically distributed).

- In MATH2089/2859/2099/CVEN2002, we can assume that the random variables in a random sampling are *i.i.d*.

Central Limit Theorem (aka the Big Man of probability)

What's this? Why do we care?

- CLT asserts:

  *For **any** random variable, the mean of a large random sample is approximately normal.*

- Basically, regardless of its original distribution, the mean will *eventually* follow a normal distribution.

Standardising the CLT

If we want to standardise the CLT...

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample taken from a population with mean μ and finite variance σ^2 and if $\bar{X}$ is the sample mean, then the limiting distribution of the standard mean follows the **standard normal distribution**. That is,

$$\frac{(\bar{X}-\mu)/\sigma}{\sqrt{n}} \overset{a}{\sim} \mathcal{N}(0,1).$$

- Note that $\overset{a}{\sim}$ means "approximately follows" (as $n \to \infty$).

Estimators

Definition: Estimators

An **estimator** of θ is a function of the sample

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n).$$

- An estimator is also a random variable!
- The most natural choice of our estimator is the sample mean! But we can have many other examples of estimators.
  - $\hat{\Theta}_1 = X_1$.
  - $\hat{\Theta}_2 = \left(\dfrac{X_1 + X_n}{2}\right)$.
  - $\hat{\Theta}_3 = \left(\dfrac{2X_1 + X_n}{2}\right)$.

Properties of estimators

Definition: Unbiased estimator

An estimator $\hat{\Theta}$ of θ is said to be **unbiased** if and only if its mean is equal to θ. That is

$$\mathbb{E}\left(\hat{\Theta}\right) = \theta.$$

- If an estimator is biased, then we can determine the bias by computing the difference

$$\text{Bias}\left(\hat{\Theta}\right) = \mathbb{E}\left(\hat{\Theta}\right) - \theta.$$

Properties of estimators

Example: Biased vs unbiased estimators

$\hat{\Theta}_1 = X_1$ is unbiased since $\mathbb{E}(\hat{\Theta}_1) = \theta$.
But $\mathbb{E}(\Theta_3) = \dfrac{1}{2}\left[2\mathbb{E}(X_1) + \mathbb{E}(X_n)\right] = \dfrac{3}{2}\theta$. So $\hat{\Theta}_3$ is biased.

Properties of estimators

Definition: Efficient estimator

Goal: An unbiased estimator should have a smaller variance. Such an estimator is said to be *more efficient*.

Example: Efficiency of estimators

$\mathrm{Var}(\Theta_1) = \sigma^2$ and $\mathrm{Var}(\Theta_2) = \dfrac{\sigma^2}{2}$. Hence Θ_2 is more efficient than Θ_1.

Properties of estimators

Definition: Consistent estimator

Goal: An unbiased estimator should also give better estimations as the number of samples grow larger. That is, an estimator is said to be *consistent* if

$$\mathrm{Var}\left(\hat{\Theta}\right) \to 0 \quad \text{as } n \to \infty.$$

Combining all three properties of estimators

We can combine all three of these properties into a single formula that tells us how accurate an estimator is. This is the **mean squared error**, which can be evaluated by computing the following

$$\mathrm{MSE}\left(\hat{\Theta}\right) = \mathrm{Var}\left(\hat{\Theta}\right) + \mathrm{Bias}\left(\hat{\Theta}\right)^2.$$

A smaller MSE means a more accurate estimator.

Part III: Confidence intervals

- Basically... we want to find a suitable range for which our estimation misses the mark with probability α. Note that α is just a percentage here!

Definition: Confidence intervals

A $100(1-\alpha)\%$ confidence interval for an unknown parameter θ is a random interval $[L, U]$, where L and U are statistics such that

$$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha.$$

- Here, our random sample has a parameter of θ!

Deriving confidence intervals

1. Find a range of values that contains $Z \sim \mathcal{N}(0,1)$ with probability $1-\alpha$.
2. Apply the result of the CLT
$$\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \stackrel{a}{\sim} \mathcal{N}(0,1).$$
3. Solve for μ for which you have a $100(1-\alpha)\%$ confidence interval for μ to be
$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Remark

If the data is exactly normally distributed, then the confidence intervals are exact!

Remark

The length of the interval measures how *precise* estimation has been! The shorter, the more precise!

Remark

Confidence intervals don't have to be symmetric! In most cases, they aren't.

Example: (MATH2089, 2018 S2 Q3bi)

In August this year, Roy Morgan Research published a poll on Rugby viewership of New Zealanders. The poll, of 6,422 randomly selected New Zealanders, found that 43.6% of them watch Rugby on the television.
Find a 95% confidence interval for the true proportion of New Zealanders who watch Rugby on the television.

Step 1.
Determine what the population proportion mean is.

$$\hat{p} = 0.436 \qquad \text{so } 1 - \hat{p} = 0.564.$$

So $SE^2 = \dfrac{0.436 \times 0.564}{6422} = 0.00003829$. So $SE = 0.006187962$.
Hence the two sided confidence interval is

$$\left[\bar{x} - z_{1-0.95/2} \times 0.006187962, \bar{x} + z_{1-0.95/2} \times 0.006187962\right].$$

Sample size determination

Margin of error

Given a pre-specified value e such that $|\bar{x} - \mu| < e$, the sample size determined is given by

$$e = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \implies n = \left(\frac{z_{1-\alpha/2\sigma}}{e} \right)^2$$

Confidence interval for a proportion

- We made some inferences about the population mean μ in the previous slides; let's move onto a population *proportion* π.

Sample proportion estimator

A useful estimator of the proportion is the **sample proportion**

$$\hat{P} = \frac{X}{n},$$

for some Binomial random variable X such that $X \sim \text{Bin}(n, \pi)$.

Sample proportion estimate

An estimate of π is simply $\hat{p} = \frac{x}{n}$.

Sampling distribution of $\hat{P}$

Applying the Central Limit Theorem to $\hat{P}$, we obtain the result

$$\frac{\hat{P} - \pi}{\sqrt{\pi(1-\pi)/n}} \overset{a}{\sim} \mathcal{N}(0, 1).$$

Additionally, we can also say that

$$\frac{\hat{P} - \pi}{\sqrt{\hat{P}(1-\hat{P})/n}} \overset{a}{\sim} \mathcal{N}(0, 1).$$

Deriving confidence intervals

1. Find a range of values that contains $Z \sim \mathcal{N}(0,1)$ with probability $1-\alpha$.
2. Apply the result of the CLT
$$\frac{\hat{P}-\pi_0}{\sqrt{\pi(1-\pi)/n}} \overset{a}{\sim} \mathcal{N}(0,1).$$
3. Solve for π for which you have a $100(1-\alpha)\%$ confidence interval for π to be
$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

One-sided confidence intervals

We can also find one-sided large-sample confidence intervals for the proportion π by finding

$$\left[0, \hat{p} + z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \quad \text{and} \quad \left[\hat{p} - z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, 1\right].$$

Part IV: Hypothesis testing

Before we begin... let's discuss an important distribution in statistics!

Student's t-distribution

A random variable T is said to follow a t_ν distribution if for $t \in \mathbb{R}$,

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

for some integer ν. Additionally, Γ is the gamma function.

- ν is the **degrees of freedom** of the distribution!

Remark

As $n \to \infty$, $t_\nu \to \mathcal{N}(0,1)$.

Null and alternative hypotheses

(Definition) Null hypothesis

For the null hypothesis H_0, we claim that our population parameter takes some sort of value.

- It is a statement that we generally believe to be true.
- We say that $H_0 : \mu = \mu_0$.

(Definition) Alternative hypothesis

For the alternative hypothesis H_1, we have some sort of "new claim" that we want to test.

- We say that $H_1 : \mu \neq \mu_0$.

Test statistic and null distribution

- To test $H_0 \mu = \mu_0$ using a random sample, when σ is known

$$Z = \frac{(\bar{X} - \mu_0)/\sigma}{\sqrt{n}} \overset{a}{\sim} \mathcal{N}(0,1).$$

- To test $H_0 : \mu = \mu_0$ using a normal random sample, when σ is not known:

$$T = \frac{(\hat{X} - \mu_0)/S}{\sqrt{n}} \sim t_\nu .$$

- To test $H_0 : \pi = \pi_0$ using a random sample

$$Z = \frac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \overset{a}{\sim} \mathcal{N}(0,1).$$

P-value

(Definition) p-values

The P-value is used to measure how much evidence there is **against** H_0 in favour of the alternative hypothesis.

The smaller the p value, the more evidence against the null hypothesis there is. If there's enough evidence against H_0, we reject the null hypothesis.

Set up of hypothesis testing

1. State the null and alternative hypotheses.
2. State the test statistic and distribution of H_0.
3. Draw a conclusion based on the corresponding p-value or rejection region.

Inferring conclusions

- At the end of the day, we want to determine whether the original claim H_0 was a lie or not. We can reach this using a **rejection region** for a statistic.
  - It is a range of values for which we would reject the null hypothesis at level α.

Hypothesis test about μ if σ is known

- Test statistic: $z = \dfrac{(\bar{x} - \mu_0)/\sigma}{\sqrt{n}}$
- Rejection region ($\mu > \mu_0$): $\left\{ \bar{x} > \mu_0 + z_{1-\alpha} \dfrac{\sigma}{\sqrt{n}} \right\}$.
- Rejection region ($\mu < \mu_0$): $\left\{ \bar{x} < \mu_0 - z_{1-\alpha} \dfrac{\sigma}{\sqrt{n}} \right\}$.
- Rejection region ($\mu \neq \mu_0$): $\bar{x} \notin \left[\mu_0 - z_{1-\alpha/2} \dfrac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2} \dfrac{\sigma}{\sqrt{n}} \right]$.

Hypothesis test about μ if σ is NOT known

- Test statistic: $t = \dfrac{(\bar{x} - \mu_0)/s}{\sqrt{n}}$
- Rejection region ($\mu > \mu_0$): $\bar{x} > \mu_0 + t_{1-\alpha,n-1}\dfrac{s}{\sqrt{n}}$.
- Rejection region ($\mu < \mu_0$): $\bar{x} < \mu_0 - t_{1-\alpha,n-1}\dfrac{s}{\sqrt{n}}$.
- Rejection region ($\mu \neq \mu_0$):
  $\bar{x} \notin \left[\mu_0 - t_{1-\alpha/2,n-1}\dfrac{s}{\sqrt{n}}, \mu_0 + t_{1-\alpha/2,n-1}\dfrac{s}{\sqrt{n}}\right]$.

Hypothesis test about π

- Test statistic: $z = \dfrac{(\bar{p} - \pi_0)}{\sqrt{\pi_0(1-\pi_0)/n}}$
- Rejection region $(\mu > \mu_0)$: $\bar{p} > \pi_0 + z_{1-\alpha}\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}$.
- Rejection region $(\mu < \mu_0)$: $\bar{p} < \pi_0 - z_{1-\alpha}\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}$.
- Rejection region $(\mu \neq \mu_0)$:
  $\bar{x} \notin \left[\pi_0 - z_{1-\alpha/2}\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}, \pi_0 + z_{1-\alpha/2}\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}\right]$.

Example: (MATH2089, 2018S2 Q3c)

Assume Rugby New Zealand (the organising body for the sport) want to be able to demonstrate that Rugby viewership is in excess of 40% of New Zealanders, using a sample of size n.
What are the appropriate null and alternative hypotheses for this test?

$$H_0 : \pi = 0.4, \qquad H_a : \pi > 0.4.$$

Example: (MATH2089, 2018S2 Q3c)

Assume Rugby New Zealand (the organising body for the sport) want to be able to demonstrate that Rugby viewership is in excess of 40% of New Zealanders, using a sample of size n.
What is the distribution of the sample proportion $\hat{p}$, if the null hypothesis is true?

$$\mathcal{N}(0.4, \sqrt{0.4(1-0.4)/n}) = \mathcal{N}(0.4, 0.4899/\sqrt{n}).$$

Assume Rugby New Zealand (the organising body for the sport) want to be able to demonstrate that Rugby viewership is in excess of 40% of New Zealanders, using a sample of size n.
Show that, for the relevant hypothesis test at the 0.05 significance level, the rejection region for $\hat{p}$ can be expressed as

$$\left(0.4 + \frac{0.806}{\sqrt{n}}, 1\right]$$

Rejection region is
$\hat{p} > \pi_0 + z_{1-\alpha}\frac{\pi_0(1-\pi_0)}{n} = 0.4 + z_{1-0.05}\sqrt{\frac{0.4 \times 0.6}{n}}$. This computes to

$$\hat{p} > 0.4 + 1.6449 \times 0.4899 \approx 0.4 + 0.806/\sqrt{n}.$$

Hence our rejection region is

$$\left(0.4 + \frac{0.806}{\sqrt{n}}, 1\right].$$

Part V: Analyses

Linear Regression

- Model the distribution of the random variable Y, conditional on the predictor X, assuming

$$Y = \beta_0 + \beta_1 x + \varepsilon.$$

The slope β_1 and the intercept β_0 are regression coefficients.

  - β_0 is the **mean** of Y when $X = 0$.
  - Slope β_1 is the change in mean of Y when X increases by 1.

Least Squares Estimators

- We often don't know the true values of β_0 and β_1. So the next best thing is to estimate them.

Notation

$$S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2$$
$$S_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

Least squares estimators of β_0 and β_1

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}, \quad \hat{\beta}_0 = \bar{Y} - \frac{S_{XY}}{S_{XX}} \bar{X}.$$

Assumptions based of the regression model

1. Conditional mean is a linear function of x. Otherwise it doesn't make any sense!
2. Each error term $e_i = y_i - (\beta_0 + \beta_1 x_i)$ are drawn independently of one another!
3. Each error term have the same variance.
4. Each error term have been drawn from a normal distribution.

Inferences about the true slope

- $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \sum_i \frac{(x_i - \bar{x})}{S_{XX}} Y_i$, where $Y \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma)$.
- Sampling distribution of $\hat{\beta}_1$ is

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma}{\sqrt{S_{XX}}}\right).$$

- Apply a hypothesis test on $\hat{\beta}_1$ with

$$H_0 : \hat{\beta}_1 = 0, \qquad H_a : \hat{\beta}_1 \neq 0.$$

- Reject H_0 if $\hat{\beta}_1$ is too different to 0. In other words, the rejection region is

$$\hat{\beta}_1 \notin \left[\hat{\beta}_1 - t_{n-2;1-\alpha/2} \frac{S}{\sqrt{S_{XX}}}, \hat{\beta}_1 + t_{n-2;1-\alpha/2} \frac{S}{\sqrt{S_{XX}}}\right].$$

Inferences about β_0

- $\hat{\beta}_0 = \sum_{i=1}^{n} \frac{Y_i}{n} - \hat{\beta}_1 \bar{x}$.
- Sampling distribution of $\hat{\beta}_1$ is

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}\right).$$

Correlation

- Recall that a regression returns a numerical relationship between two random variables. On the other hand, a correlation quantifies the strength of the linear relationship between X and Y. We can show that the sample correlation coefficient is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

Analysis of Variance (ANOVA)

- We use analysis of variance when dealing with k random samples, where $\bar{X}_i$ and S_i are the sample mean and standard deviation of the ith sample.

ANOVA model

$$X_{ij} = \mu_i + \varepsilon_{ij},$$

where μ_i is the mean at the ith treatment and ε_{ij} is an individual random error component.

Assumptions

$$\varepsilon_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma).$$

- Errors are normally distributed, are independent and have the same variance.

ANOVA hypotheses

- Null hypothesis: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.
- Alternative hypothesis: H_a : not all means are the same.
    - We're not saying that ALL means are different, but that at least two means are different.

Fisher's F-distribution

Let $f_{d_1,d_2;\alpha}$ be a value such that

$$\mathbb{P}(X > f_{d_1,d_2;\alpha}) = 1 - \alpha,$$

where X follows an F_{d_1,d_2} distribution with density

$$f(X) = \frac{\Gamma((d_1+d_2)/2)(d_1/d_2)^{d_1/2}x^{d_1/2-1}}{\Gamma(d_1/2)\Gamma(d_2/2)((d_1/d_2)x+1)^{(d_1+d_2)/2}}.$$

Yeah nah, I don't remember this at all! They would normally give you a value by computing the command `finv(`α`, d1, d2)` for quantiles and `1-fcdf(x,d1,d2)`.

ANOVA test

- Use the test statistic
$$f = \frac{\mathrm{ms}_{\mathrm{Tr}}}{\mathrm{ms}_{\mathrm{Er}}},$$
where f follows a Fisher distribution with $d_1 = k - 1$ and $d_2 = n - k$.
- Reject H_0 if
$$\frac{\mathrm{ms}_{\mathrm{Tr}}}{\mathrm{ms}_{\mathrm{Er}}} > f_{k-1,n-k;1-\alpha},$$
where $\mathrm{ms}_{\mathrm{Tr}}$ is the treatment mean squared and $\mathrm{ms}_{\mathrm{Er}}$ is the mean squared error.