Week 5: Statistics 2

Statistical Inference and Hypothesis Testing

Dr Giuseppe Brandi

Northeastern University London

Outline

- Definition
- Dependence vs Independence Conditional Probability
- Bayes's Theorem, Populations and Random Samples
- Random Variables, Probability Density Function, Probability Distribution
- Hypothesis Testing, z-test, t-test
- Dependent Sample t-test, Independent t-test

Probability

What is a probability?

Probability is a way of quantifying the uncertainty associated with events chosen from some universe of events.

Consider rolling a die: it consists of all possible outcomes. Any subset of these outcomes is an event; for example, "the die rolls a one" or "the die rolls an even number."

Probability: Dependence vs Independence

We say that two events E and F are dependent if knowing something about whether E happens gives us information about whether F happens (and vice versa).

Otherwise, they are independent.

Probability: Conditional Probability

The probability of event E given that event F has occurred is called the conditional probability of E given F:

$$P(E|F) = \frac{P(E \text{ and } F)}{P(F)}$$

Probability: Bayes's Theorem

It is a way of "reversing" conditional Probability. Assume we need to know the probability of some event E conditional on some other event F occurring. But we only have information about the probability of F conditional on E occurring.

Populations & random samples

Take a random sample from a population, P

Use sample to make inference about P

In data science and statistics we make use of sample data to infer population information.

Random variables

Definition:

A random variable is a variable whose possible values have an associated probability distribution.

Example: A very simple random variable equals 1 if a coin flip turns up heads and 0 if the flip turns up tails. A more complicated one might measure the number of heads you observe when flipping a coin 10 times or a value picked from range(10) where each number is equally likely.

Random variables

The associated distribution gives the Probability that the variable realizes each of its possible values.

Example: The coin flip variable equals 0 with probability 0.5 and 1 with probability 0.5.

The range(10) variable has a distribution that assigns probability 0.1 to each of the numbers from 0 to 9.

Random variables

Expected value of a random variable is the average of its values weighted by their Probability.

Example: The coin flip variable has an expected value of $1/2 = 0 * 1/2 + 1 * 1/2$.

Range(10) variable has an expected value of 4.5.

Discrete random variables

A random variable X has a set function that assigns one real number to each possible outcome (the sample space, S) For example, rolling a dice, X = 1, 2, 3, 4, 5, 6 We can then ask, P(X = 1)? There are discrete and continuous random variables

A discrete random variable X has either a finite or countable number of possible outcomes of X.

For example, consider rolling a die:

$X = 1$ or $X = 2$, and so on; or

X equals the number of rolls until the dice lands on 6: X = 1, 2, 3, ...

Continuous random variables

A continuous random variable X has an infinite number of possible outcomes of X.

Examples include:

- X is "the weight of a randomly selected person."

- X is "the height of a randomly selected tree."

A probability density function (p.d.f.) of a continuous random variable X is a function $f(x)$:

$f(x)$ is positive everywhere, and the area under the curve is 1.

Probability Distributions

Density estimation problem:

Given a finite set $x_1, x_2, \ldots, x_n$ of observations, model the probability distribution $p(x)$ of random variable x.

Observations are independent & identically distributed (iid). There are infinitely many $p(x)$ as candidates.

So, the problem to consider is model selection.

Uniform distribution

A continuous random variable X has a uniform distribution $U(a, b)$ if the probability density function (p.d.f.) is:

$$f(x) = \frac{1}{b-a}$$

for constants a and b such that $a < x < b$.

Uniform distribution

Although not representative of randomness seen in real world, $U(a, b)$ is useful, say, for pseudo-random number generation.

"Select a random individual from population" Set random seed:

```
import numpy as np
Ensure reproducibility of RNG
np.random.seed(123456789)
```

Generate a Gaussian random sample of 100 numbers from 0 to 1

```
import numpy as np
samples = np.random.uniform(0, 1, 100)
assert np.all(samples >= 0)
assert np.all(samples < 1)
```

Uniform distribution: discrete choices

Generate a uniform random sample of size 10 from space A–D:

```
import numpy as np
np.random.choice(['A', 'B', 'C', 'D'], 10)
```

```
Result:[D, B, D, D, A, B, A, B, C, C]
```

Normal (Gaussian) distribution

The normal distribution $N(\mu, \sigma^2)$ is the most prevalent probability distribution in the natural world. It has a characteristic bell shape.

The bell shape depends on the mean μ and variance σ^2 (or standard deviation, σ).

Normal distribution: random floats

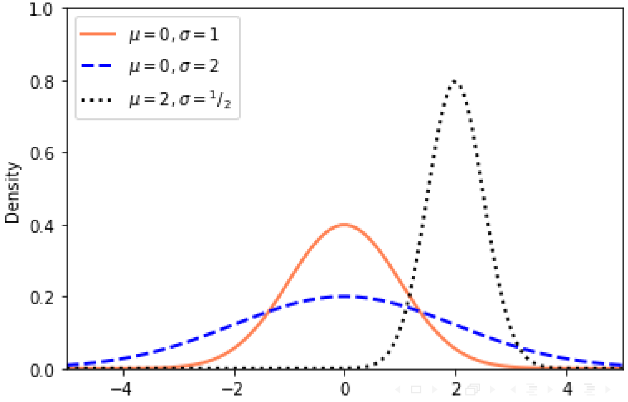
Generate a uniform random sample of 100 numbers with mean 100 and standard deviation 16:

Python Code

```
import numpy as np
samples = np.random.normal(loc=100., scale=16., size=100)
```

Normal (Gaussian) distribution

The bell shape depends on the mean μ and variance σ^2 (or standard deviation, σ).



Normal (Gaussian) distribution

The bell shape depends on the mean μ and variance σ^2:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Normal (Gaussian) distribution

Compute p.d.f. of standard normal distribution:

Python Code

```
import numpy as np
from scipy.stats import norm
x = np.random.uniform(-5., 5., 100)
y = norm.pdf(x, 0.0, 1.0)
plt.scatter(x, y)
```

Normal distribution: random floats

Generate a uniform random sample of 100 numbers with mean 100 and standard deviation 16:

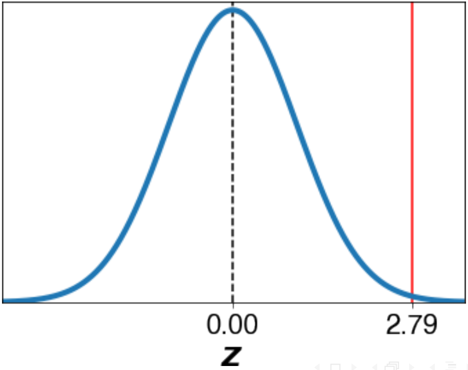
Python Code

```
samples = np.random.normal(loc=100., scale=16., size=100)
```

Note: We could generate a histogram of samples.

z-distribution

Problem: ''Suppose $X \sim N(100, 16^2)$ is the IQ of a random person. What is $P(X \leq 90)$? It is not possible to compute the area under the curve of a normal p.d.f.



z-distribution

However, if $X \sim N(\mu, \sigma^2)$, then:

$$Z = \frac{X - \mu}{\sigma}$$

Follows the standard normal distribution $N(0, 1)$.

Hypothesis testing

We want to test whether a certain hypothesis is likely to be true. Hypotheses are assertions. Examples:

- "This coin is fair."

- "Data scientists prefer Python to R."

Null hypothesis H_0 represents some default position. Alternative hypothesis H_1 is what we compare it with.

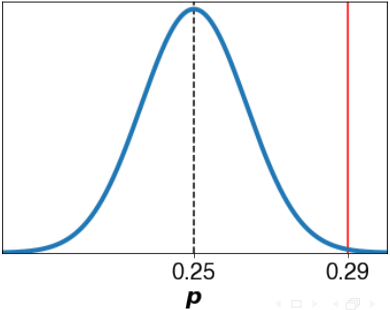
Hypothesis testing in 3 steps

The basic procedure is:

1. State an initial (or null) hypothesis about the parameter, H_0.
2. Collect evidence (in the form of data).
3. Based on data, reject or not the null hypothesis H_0.

Hypothesis testing: an example

Suppose that a dealer draws (with replacement) 1,000 cards from an (infinite) deck and 290 of those cards are hearts. Is the card dealer fair? i.e., $H_0 : p_0 = 0.25$ (and $H_1 : p_0 > 0.25$). Alternatively, a z-test statistic should be used (see next slides). There are 4 suits: hearts, spades, clubs, and diamonds.



Hypothesis testing: z-test

What is the Z-Test Statistic?

z-tests are a statistical way of testing a Null Hypothesis when either:

- We know the population variance, or

- We do not know the population variance, but our sample size is large $n \geq 30$.

Hypothesis testing: z-test

z-test vs t-test:

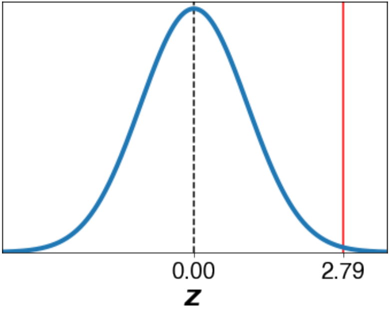
If we have a sample size of less than 30 and do not know the population variance, we must use a t-test. It is assumed that the z-statistic follows a standard normal distribution. But the t-statistics follow the t-distribution with degrees of freedom equal to $n-1$, where n is the sample size.

Note: The samples used for z-test or t-test must be independent samples and must have a distribution identical to the population distribution.

Hypothesis testing: z-test

A test statistic, z, follows a normal distribution $N(0,1)$:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$



Hypothesis testing: z-test - Example

Computing Z-test and P-value:

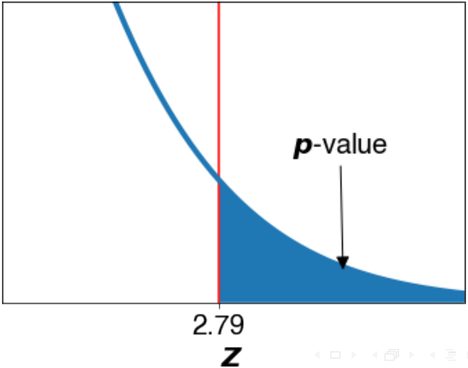
```
from statsmodels.stats
import proportion as pr
Z, P = pr.proportions_ztest(290, 1000, 0.25,
alternative="larger")
print(f"Z = Z:.2f, P-value is P:.4f")
```

```
Result:  Z = 2.79, P-value = 0.0027
```

Hypothesis testing: z-test - Example: What is the p-value?

The P-value is the area under the curve. The smallest significant level leads to the rejection of the null hypothesis H_0.

We say, "If $P \leq \alpha$, then reject H_0." Typical values for α are 0.01, 0.05, and 0.10.



Hypothesis testing: z-test - Example: Possible Errors

Type I Error: We reject H_0 (in favour of H_1) when in fact H_0 is true. α translates into our willingness to commit a Type I Error.

Type II Error: We accept H_0 when in fact H_0 is false.

Hypothesis testing: t-test

What is the t-test?

t-tests are a statistical way of testing a hypothesis when:

- We do not know the population variance, or

- Our sample size is small $n < 30$.

Hypothesis testing: t-test

Hypothesis tests based on the t-distribution for one or more population means, assuming their variance is unknown:

- Test the mean μ of a single population.

- Compare means μ_X and μ_Y of two dependent populations X and Y.

- Compare means μ_X and μ_Y of independent X and Y.

Hypothesis testing: t-test - Example

We perform a **One-Sample** t-test when we want to compare a sample mean with the population mean.

We perform a **Two-Sample** t-test when we want to compare the mean of two samples.

The difference from the z-test is that we do not have information on the population variance here.

We use the sample standard deviation instead of the population standard deviation in this case.

Hypothesis testing: t-test Example

Suppose a fast-food chain claims its burger weighs 113g. A customer sampled 100 burgers and found an average weight of 110g with a standard deviation of 19.4g.

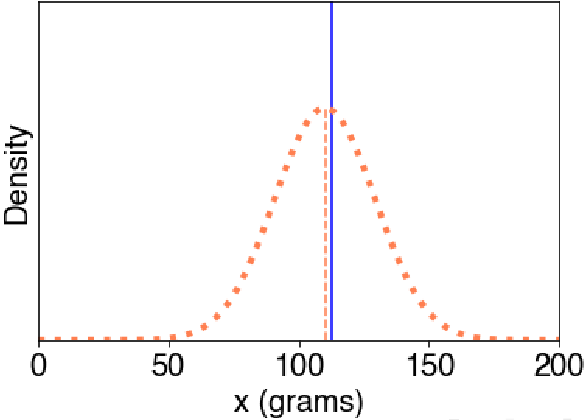
Python Code

```
a = np.random.normal(loc=110, scale=19.4,
size=100)
print(f"a.mean():.2f, a.std():.2f")
```

```
Result:  110.10, 19.39
```

Hypothesis testing t-test: Example

Suppose that $H_0 : \mu_0 = 113$ (and $H_1 : \mu_0 \neq 113$).



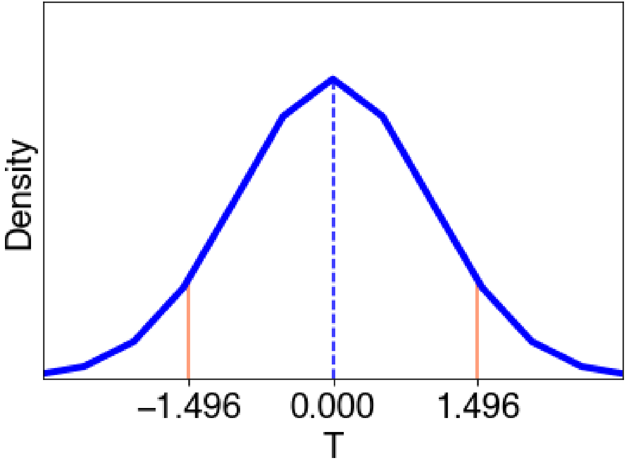
Hypothesis testing t-test: Example

If $X \sim N(\mu, \sigma^2)$, then the t-test is computed by:

$$t = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

follows a t-student distribution with $n - 1$ degrees of freedom.

Hypothesis testing t-test: Example



Hypothesis testing t-test: Example

t-test:

Python Code

```
from scipy.stats
import ttest_1samp
r = ttest_1samp(a, 113)
print(f't=r.statistic:.2f, p=r.pvalue:.2f')
```

```
Result:  t = -1.49, p = 0.14
```

The P-value is greater than $\alpha = 0.01$, so we cannot reject H_0.

Hypothesis testing: Paired t-test

Paired t-test:

It is a statistical procedure used to determine whether the mean difference between two sets of observations is zero.

In a paired sample t-test, each subject or entity is measured twice, resulting in pairs of observations.

Common applications include case-control studies or repeated-measures designs.

Paired t-test: Example

Assume you are interested in evaluating the effectiveness of a company training program.

One approach to consider would be to measure the performance of a sample of employees before and after completing the program, and analyze the differences using a paired sample t-test.

Hypothesis testing: Two-sample t-test

Two-sample t-test:

A two-sample independent t-test can be run on sample data from a normally distributed numerical outcome variable to determine if its mean differs across two independent groups.

Two-sample t-test: Example

We could apply a two-sample independent t-test to find out whether the mean GPA differs between freshman and senior college students by collecting a sample of each group of students and recording their GPAs.

Hypothesis testing: Paired t-test - Python Example


```python
from scipy.stats import ttest_rel
import numpy as np

# Simulate data for before and after training program
before = np.random.normal(loc=50, scale=10, size=30)
after = np.random.normal(loc=55, scale=10, size=30)
# Paired t-test
t_stat, p_value = ttest_rel(before, after)

print(f"t = {t_stat:.2f}, p-value = {p_value:.4f}")
```

```
t = -1.18, p-value = 0.2484
```

Hypothesis testing: Two-sample t-test


```python
from scipy.stats import ttest_ind
import numpy as np

# Simulate GPA for two groups of students:
# freshmen and seniors
freshmen_gpa = np.random.normal(loc=2.8, scale=0.3, size=50)
seniors_gpa = np.random.normal(loc=3.2, scale=0.4, size=50)
# Two-sample independent t-test
t_stat, p_value = ttest_ind(freshmen_gpa, seniors_gpa)

print(f"t = {t_stat:.2f}, p-value = {p_value:.4f}")
```

```
t = -4.15, p-value = 0.0001
```

Summary

- Probability: Definition

- Probability: Dependence vs Independence

- Conditional Probability

- Bayes's Theorem

- Populations and Random Samples

- Hypothesis Testing: z-test, t-test, Dependent t-test, Independent t-test.