The University of Texas at Austin
School of Information

Stats: Interview Questions

Mick McQuaid

2024-04-29

LAST CLASS

Statistical significance

How would you assess the statistical significance of an insight?

Statistical significance

Perform hypothesis testing. State the null and alternative hypothesis, calculate the p-value, then set the level of significance desired, α. If $p < \alpha$, claim that the result is statistically significant.

Central Limit Theorem

What is it? Can you explain it? Why is it important?

Central Limit Theorem

The sampling distribution of the sample mean approaches a normal distribution as the sample size grows, no matter the shape of the population distribution.

It's important any time we compare means or calculate confidence intervals of means.

Statistical power

What is it?

Statistical power

Statistical power is the ability of a test to detect a difference between two quantities. A more formal definition is that power is the probability that you reject the null hypothesis if the alternative hypothesis is true. In practice, people typically strive for a statistical power of at least 0.8, but this can vary depending on the criticality of the result.

Selection bias

What is selection bias with respect to a data frame?

Selection bias

Selection bias occurs when you don’t take a random sample. This can be done in many ways, including nonrepresentative time intervals, attrition, replacing missing data with the mean of the data, susceptibility to illness, cherry-picking, suppressing evidence, and stopping data collection as soon as a desired result is achieved.

Experimental vs observational data

What is the difference between experimental and observational data?

Experimental vs observational data

Experimental data is collected under controlled conditions, while observational data is collected outside the laboratory and may be affected by uncontrolled factors. Working with observational data is often called quasi-experimentation.

Imputation of missing data

How should you deal with missing data? Should you replace it by the mean?

Imputation of missing data

You should never replace missing data with its mean because that reduces the variance of the data. This crude technique doesn’t take feature correlation into account. For better ways to flexibly impute missing data, see van Buuren (2018).

Outliers

What is an outlier and how might you handle it?

Outliers

An outlier is a point that differs significantly from other observations. You should always check the reasons for outliers to determine whether they were produced by a process that inevitably has outliers, in which case you should include them in your analysis, or some other reason, in which case you should omit them. You can identify them with boxplots or regression diagnostic plots.

Call duration

Suppose you have the duration of calls to a call center. How would you analyze this data?

Call duration

First, visualize the data with a histogram. At the low end, the data is bound by 0 seconds, so you might expect a lognormal distribution, that is it will be positively skewed, with a few lengthy calls. Use a QQ plot to determine whether it actually does follow a lognormal distribution. If you have a small number of agents, it may be easy to make a QQ plot for each one.

Probability of rain

Suppose you plan to visit a place where it generally rains one quarter of the time. To decide whether to bring an umbrella, you call three people there, who all tell you it is raining. Each one has a two thirds probability of telling the truth and a one third probability of lying. What is the probability of rain?

Probability of rain

This is a typical Bayes rule problem, where we have r is the event that it is raining and c is the event that all three claim it is raining. So $p(r) = 0.25$ and $p(c|r) = (2/3)^3 = 8/27$. We would like to know $p(c)$ and $p(r|c)$.

$$p(c) = p(c|r)p(r) + p(c|\neg r)p(\neg r)$$

$$p(c) = (2/3)^3 * (1/4) + (1/3)^3 * (3/4) \approx 0.1$$

$$p(r|c) = p(c|r)p(r)/p(c) = (1/4) * (8/27)/0.1 \approx 0.74$$

There's slightly less than a three in four chance it's raining.

Cards

Consider two card decks, one with 12 black and 12 red cards, and one with 24 black and 24 red cards. Suppose you draw two cards randomly from a deck. Which deck has the greater probability of two cards of the same color?

Cards

Suppose you draw a red from the first deck. The probability of drawing a second red is $11/(11 + 12) \approx 0.478$. On the other hand, if you draw a red card from the second deck, the probability of another is $23/(23 + 24) \approx 0.489$!

Non-normal distributions

What are examples of non-normal and / or non-lognormal distributions?

Non-normal distributions

Any categorical data has a non-normal distribution, as does data with an exponential distribution.

Median vs mean

When is the median a better measure of centrality than the mean?

Median vs mean

The median dominates the mean whenever the mean is skewed, such as by outliers.

Fair dice

What is the probability of rolling two dice with a sum of four?

Fair dice

The probability is $3/36 \approx 0.083$

Law of Large Numbers

What is the law of large numbers?

Law of Large Numbers

The law of large numbers states that, as the number of trials increase, the average converges to the expected value.

Margin of error

What is the margin of error and how do you calculate it?

Margin of error

The margin of error tells the amount of random sampling error in a sample, such as a survey. It is calculated as

$$MOE_{\gamma} = z_{\gamma}\sqrt{\frac{\sigma^2}{n}}$$

Sampling bias

What are some examples of sampling bias?

Sampling bias

Examples include non-random sampling, sampling too few observations, and survivorship bias (overlooking observations that were unsuccessful in a multistage selection process)

Controlling bias

How do you control bias?

Controlling bias

Ensure that each observation has an equal probability of selection.

Confounding variables

What is a confounding variable?

Confounding variables

A variable is confounding when it influences both a dependent and independent variable, leading to a non-causal relationship between the dependent and independent variable. A classic example is shark sightings and ice cream sales at the beach, where temperature is the confounding variable.

A/B Testing

What is A/B testing?

A/B Testing

Any two sample hypothesis test with a control sample and a variant can be construed as an A/B test. Ordinarily, the term is applied to two slightly different webpages with some sort of conversion (e.g., sale) as an outcome variable.

Infection rate

Suppose infection rates above 1 per 100 person-days at risk are considered high. A given hospital had 10 infections over 1,787 person-days at risk. Provide the p-value for the correct one-sided test of whether the hospital has failed the standard.

Infection rate

`ppois(10,1787*1/100)` is the correct formula to use in R. It has a value of 0.0323715.

Biased coin

Suppose a coin is biased so that $p(\text{heads}) = 4/5$. *What is the probability of getting three or more heads in five flips?*

Biased coin

$$p(5h) + p(4h) + p(3h) = (4/5)^5 + (4/5)^4 * (1/5)^1 + (4/5)^3 * (1/5)^2 \approx 0.94$$

Random variable

Let X be a normally distributed random variable with mean 1020 and standard deviation 50. What is the probability that $X > 1200$?

Random variable

The correct R formula would be `1-pnorm(1200,1020,50)` and its value in this case is 1.5910859^{-4}.

Customers calling

Suppose that, on average, 2.5 customers call a call center every minute. What are the chances that at most 3 people call in four minutes?

Customers calling

This is a standard Poisson distribution problem, also framed as customers arriving or buses arriving or customers joining a queue. The correct R formula is `ppois(3,2.5*4)` and its value is 0.0103361.

HIV testing

Suppose an HIV test has a sensitivity of 99.7% and a specificity of 98.5%. Suppose the general population has an infection rate of 0.001 and particular person receives a positive result. What is the precision of the test or, in other words, what is the probability that this person is HIV positive?

HIV testing

Let’s use a formula from Wikipedia for precision based on sensitivity, specificity, and prevalence for positive predictive value (PPV) which is a measure of precision:

$$\begin{aligned}
\text{PPV} &= \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})} \\
&= \frac{0.997 * 0.001}{0.997 * 0.001 + (1 - 0.985) * (1 - 0.001)} \\
&\approx 0.06238268
\end{aligned}$$

Polling

Your pollster polled 100 people before a two person election and sixty said they'd vote for you. What is a 95% confidence interval for this statistic?

Polling

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$.6 \pm 1.96\sqrt{.6(1-.6)/100}$$

$$\approx [50.4, 69.6]$$

References

van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press.

Colophon

This slideshow was produced using `quarto`

Fonts are *Roboto Condensed Bold*, *JetBrains Mono Nerd Font*, and *STIX2*