Mathematical statistics

Chris Johnson

12/4/22

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Distributions

- `d` is $\Pr(X = x)$ (PDF)
- `p` is $\Pr(X \leq x)$ (CDF)
- `q` is quantile function (similar to PDF)
- `r` is random sample

2 Identification

Quantile-Quantile (Q-Q) plots can be formed for any distribution

Quantiles are lines that divide the data into equally-sized groups Describes the amount of data to the left (or below) it

- Calculate the quantiles for each n observations in the sample (0–100)
- Divide the distribution into n equally-probable parts ($-$ inf–inf)
- Plot the data quantile against the distribution quantile
- If the points adhere to $y = ax + b$ (depends on the distribution), then this provides evidence the data follow the distribution

2.1 Families

3 Distribution relationships

3.1 Table of relationships

Distribution	Relatives	Notes
Poisson	Binomial	Can be used to approximate the Binomial
Negative Binomial	Pascal	Special case of Pascal

3.2 Family

- Log-normal
- Normal
- t

3.3 Family

Pascal Negative binomial

3.4 Family

Beta Gamma Exponential

3.5 Identification

Keywords can be used to nail down the distribution.

Keyword	Distribution	Notes
Non-conforming items	Hypergeometric	
Successes, failures	Binomial	Can be approximated by the Poisson

Keyword	Distribution	Notes
Per unit	Poisson	
Pascal		
Time-to-failure	Exponential	

3.6 Discrete distributions

Random variables are vectors.

The probability mass function (PMF) $p(x)$ is any function that

3.6.1 Poisson

Rate

The Poisson distribution is appropriate when calculating rate, the number of events per unit. Note that unit can be spatial (number of purse snatchings per 10 quadrats) or temporal (number of phone calls per 22 minutes).

As $lambda \to \infty$, the Poisson distribution approximates the normal distribution

3.6.2 Binomial

Binomial distribution has two parameters:

- number of trials
- probability of success

Observation refers to observing one or more trials. Multiple observations means observing one or more trials multiple times. Example: Your daredevil cat goes over Niagara Falls in a barrel, and either survives or dies. (Cats have nine lives.)

One observation of three trials means you go over Niagara Falls three times. Each time, the outcome is recorded: survive is recorded as success; death is recorded as failure.

Two observations of three trials means you simply repeat this experiment twice, resulting in two datasets.

The trials are independent: for all trials, the outcome of any single trial doesn't influence the outcome of the other trials. (Feelings don't matter to the barrel or the waterfall.)

Situations that are not naturally binomial may be able to be dichotomized, assuming the trials are independent.

Example: A combination lock has five dials, and each dial has four letters: A, B, C, and D. For the lock to open, the dial must be turned to the correct letter. The dial can be either correctly set or incorrectly set, which corresponds to success or failure respectively. Additionally, since the outcome of setting of one dial doesn't influence the outcome of setting any of the other dials, the dials can be treated as trials, and those trials can be assumed independent.

If a crook has one chance to guess the correct combination, the probability that the crook sets two of the five dials to the correct position is 0.2636719.

The probability of success is 0.25. This is because without any knowledge of the correct position of the dial, each of the four letters are equally likely, hence $0.25 = \frac{1}{4}$.

In R,

```
dbinom(x = 2, size = 5, prob = 0.25)
```

`n` is the number of observations.

Note, `sum(dbinom(x = 0:5, size = 5, prob = 0.5))` is 1.

The probability that the crook gets less than five correct is 0.9990234, which is equivalent to

```
sum(
  dbinom(
    x = 0:4,
    size = 5,
    prob = 0.25
  )
)
```

or

```
pbinom(q = 4, size = 5, prob = 0.25)
```

`pbinom()` gives $\Pr(X \leq x)$. `pbinom()` uses the cumulative

Bin(trials, probability of success)

3.6.3 Poisson

Discrete

bounds: $[0, \infty)$

$\lambda \in \mathbb{R}$

Describes number of events occurring in a fixed time interval or region of opportunity

λ is the expected number of events.

Assumptions:

- rate is constant
- events are independent (no events influence other events)

Probability mass function:

$\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$

Cumulative distribution function:

$\Pr(X \leq x) = \frac{\Gamma(x+1\,\lambda)}{\lfloor x! \rfloor}$

$\mathbb{E}(X) = \lambda = \mathbb{V}(X)$

3.6.4 Hypergeometric

- N total items
- D non-conforming items
- Wish to sample n items

$$p(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$$

Probability of $x \in \{0, 1, 2, \ldots, \min(n, D)\}$ non-conforming samples:

$$\Pr(X \leq x) = p(0) + \ldots + p(x)$$

3.6.5 Binomial

- Independent Bernoulli trials (success or failure)
- Quality engineering
- Can be used to approximate the hypergeometric
- p is the fraction of non-conforming items
- n is the sample size
- x is the number of successes
  - Success can be a defect, e.g. a non-conforming item
  - Want to know number of non-conforming items in the sample

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\Pr(\hat{p} \leq a) = \Pr\left(\frac{x}{n} \leq a\right) = \Pr(x \leq na)$$

- Sample size is fixed
- Count the number of successes

Note: Time to review changing the limits of summation and integration!

3.6.6 Poisson

- Phenomenon occurring on a per-unit basis
  - defect
  - per unit area, length, time, etc.
- Can be used to approximate the binomial

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

3.6.7 Pascal

- Bernoulli trials
- p is the probability of success
- r is the number of successes
- x is the trial

This distribution gives the probability that the rth success occurs on trial x, based on a probability of success:

$$p(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

3.6.7.1 Negative binomial

- Special case of the Pascal
- Fix the number of successes
- Used to determine the sample size needed to observe x successes
- $r \in \mathbb{R}$ and $r > 0$

3.6.7.2 Geometric

- Special case of the Pascal, when $r = 1$.

3.7 Continuous distributions

Random variables are intervals.

The probability density function (PDF) $f(x)$ is any function that produces a curve with the following two properties $\forall x$:

- $f(x) \geq 0$
- $\int_a^b f(x)\mathrm{d}x = 1$

The curve is called the density curve, and the area is referred to as the density. Since the total area under the density curve (or equivalently, the area of the density) is equal to 1, the values can be interpreted as probabilities.

The density curve is said to be supported over an interval. Support describes the interval. Discrete distributions either have finite or infinite support. Continuous distributions either have bounded ($[a, b]$), semi-infinite ($[0, +\infty)$), or infinite support ($(-\infty, +\infty)$).

The cumulative distribution function (cdf) is the sum of the area under the density curve from the minimum supported value to a value of interest, giving $\Pr(X \leq x)$, where X is the random variable (RV) and x is the value of interest the RV can take on. It follows that

$$\Pr(X \leq x) + \Pr(X > x) = 1$$

and by arithmetic

$$\Pr(X > x) = 1 - \Pr(X \leq x)$$

$\Pr(a \leq x \leq b)$ is calculated using the cdf:

$$\Pr(a \leq x \leq b) = \Pr(X \leq b) - \Pr(X \leq a)$$

For continuous distributions, $\Pr(X = x) = 0$. Let $a = b$:

$$\begin{aligned}\Pr(a \leq x \leq b) &= \Pr(X \leq b) - \Pr(X \leq a) \\ &= \Pr(X \leq b) - \Pr(X \leq b) \\ &= 0\end{aligned}$$

3.7.1 Normal

- Can be used to approximate the binomial
-

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

3.7.2 Lognormal

3.7.3 Exponential

- Used in reliability engineering (time to failure)
  - λ is called the failure rate
  - $\frac{1}{\lambda^2}$ is called the mean time to failure

$$f(x) = \lambda e^{-\lambda x}$$

See note about relation to the Poisson distribution on p. 70 of Montgomery.

3.7.4 Gamma

- Becomes the exponential distribution when $r = 1$
- Parameters shape and scale the distribution

$$f(x) = \frac{\lambda}{\Gamma(r)}(\lambda x)^{r-1}e^{-\lambda x}$$

3.7.5 Weibull

- Used in reliability engineering
  - Time to failure
- Reduces to the exponential distribution when $\beta = 1$

$$f(x) = \frac{\beta}{\theta}\left(\frac{x}{\theta}\right)^{\beta-1}\exp\left[-\left(\frac{x}{\theta}\right)^{\beta}\right]$$

4 Fundamentals

4.1 Correlation

covariance: direction, not standardized correlation: strength and direction, standardized, function of covariance

4.2 Counting

Suppose we have three urns. All urns have numbered balls.

The first urn can hold two balls. The second urn can hold five balls. The third urn can hold ten balls.

Suppose we ask someone to choose

- one ball from the first urn
- three balls from the second urn
- three balls from the third urn

How many outcomes are there?

The first urn has two balls, and we're drawing one:

There are $\binom{2}{1=2}$ outcomes for the first urn.

The second urn has five balls, and we're drawing three:

There are $\binom{5}{3=10}$ outcomes for the second urn.

The third urn has ten balls, and we're drawing three:

There are $\binom{10}{3=120}$ outcomes for the third urn.

Therefore, the total number of outcomes for choosing seven balls from the three urns, when we must choose one of seven from the first urn, three of seven from the second urn, and three of seven from the third urn is $2 \times 10 \times 120 = 2400$.

4.3 Calculus

4.4 Linear algebra

5 Quantitative

quantitative	inference
1	μ
2	ρ, regression

5.1 Categorical

variables	categories	inference
1	2	p, χ^2 goodness of fit
1	3 or more	χ^2, goodness of fit
2	2	$ p_a - p_0$ χ^2, test for association
2	3 or more	χ^2, test for association

5.2 Quantitative and categorical

categories	inference
2	$\mu_0 - \mu_a$, ANOVA
3 or more	ANOVA

5.3 Regression

response	inference
quantitative	regression
categorical	logistic regression

5.4 Pitfalls

5.4.1 Multiple testing

A Type-I error occurs when the null hypothesis is rejected when it is actually true. The chance that a Type-I error occurs increases as more hypothesis tests are conducted. $100-100(1-\alpha)\%$ of tests will yield statistically significant results by chance.

- 7129 genes
- 38 patients
- Each patient has one of two types of leukemia
- Each gene was tested for a difference between two types of leukemia

Assume there are no genetic differences between the leukemias. If $\alpha = 0.01$, it can be expected that 1% of the 7129 tests (71 or 72) to yield statistically significant results by chance.

To avoid this pitfall, divide α by the number of tests. E.g., since there are 7129 tests, it can be expected that practically no tests will yield statistically significant results by chance. (*Technically* 0.01 tests.)

6 Regression

Extension of the inference chapter.

Dependent variable	Regression model
Continuous	Linear
Binary	Logistic
Multicategory (nominal)	Multinomial logit
Multicategory (ordinal)	Cumulative logit
Count	Poisson

If all explanatory variables are categorical, we model contingency tables.

H_0: Equiprobable model (assume uniform distribution; expected value is inverse of number of categories)

Test if observed is different than expected.

Goodness of fit statistics for Poisson regression.

Pearson χ^2 test Deviance or log-likelihood ratio test for Poisson

$\chi^2 = \sum_i^n \left[\frac{(O_i - E_i)}{sqrt(E_i)}\right]^2$

$L^2 = 2\sum_i^n O_i \log\left(\frac{O_i}{E_i}\right)$

Both should follow χ^2 with d.f. = number of cells - number of model parameters

Both χ^2 and L^2 are asymptoticaly equivalent (as $n \to$ inf, both converge to χ^2; both rely on large samples)

Diagnostics

Residual analysis

7 Sampling

Squid

Experiment: Ensure flow rate is 525 ± 25 mL for some number of trials.

Task, stated in many ways:

Range: [500, 550]

Calculate the number of trials needed to ensure the flow rate 525 ± 25

Sample size, 15, 39 choice of, 124-129 effect on chi-squared statistic, 235 effect on width of confidence interval, 114, 133

7.1 15

Simple random sample Each possible sample of size n has the same probability of being selected.

Section 2.4 presents more complex sampling schemes.

7.2 39

No useful information.

7.3 124-129

Margin of error

The margin of error for a confidence interval depends on the standard error of the point estimate. See Exercise 57 in Chapter 4. Complex sampling schemes mentioned are stratification and clustering. Standard errors for the complex sampling schemes are approximated by the formulas for SRSs, either as-is or inflated by a factor

- The margin of error depends directly on the standard error of the sampling distribution of the point estimator.
- The standard error itself depends on the sample size.

To determine the sample size we must

- decide on the margin of error desired
- specify the probability with which the margin of error is achieved

Example 5.6 Sample Size for a Survey on Single-Parent Children

Determine n such that a $100 \cdot (1-\alpha)\%$ confidence interval for θ equals $\hat{\theta} \pm \text{MOE}$.

Sampling distribution :

If the sampling distribution of $\hat{\theta}$ is approximately normal, $\hat{\theta}$ falls within 1.96 standard errors of θ with probability 0.95.

Determine $\sigma_{\hat{\theta}}$.

If $\theta = \pi$, then $\sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}}$. If $\theta = \mu$, then ? If $\theta = \tau$, then ?

Regardless,

Margin of error equals test statistic times true standard error.

MOE = ?

7.4 235

7.5 114

7.6 133

Sampling scheme	In general
Simple	Use this tutorial
Stratified	Need less, due to less variability in subpopulations than the population
Cluster	Need more
Multistage	Need more

Sample size depends on

1. margin of error (precision)
2. probability the confidence interval will contain the parameter (confidence)
3. variability in the population
4. complexity of analysis
5. resources (time, money, etc.)

With probability $100 \cdot (1 - \alpha)\%$, the sample size needed to estimate correctly within .

If normal,

k standard deviations	Probability
1	0.680
2	0.950
3	0.999

Point estimate Interval estimate

If you need to guess $\text{Var}\theta$, use the fact that the range is $2 \cdot k\text{Var}\theta$, then

1. establish a reasonable range
2. substitute the range for $\sqrt{\text{Var}\theta}$ in $2 \cdot k\sqrt{\text{Var}\theta}$
3. solve for $\text{Var}\theta$

H_0 false	H_0 true
Reject H_0	Power $(1-\beta)$	Type I error with probability α
FTR H_0	Type II error with probability β

For two-tailed situations, use $\frac{\alpha}{2}$ when finding the z-score.

$n = \sqrt{\text{Var}\theta} \cdot \left(\frac{z}{MOE}\right)^2$

7.7 What to do when there is no way around a small sample size

Use a t method Look for extreme outliers great departures from normal population assumption (tests usually use the mean, and assume the location of the mean is the center of the distribution; if the distribution is skewed, the mean is not the center.)

8 Theorems

8.1 The Central Limit Theorem

SE for sample proportion: SE $= \sqrt{\frac{p(1-p)}{n}}$. (As n increases, $\frac{p(1-p)}{n}$ decreases.) See derivation of formula. SE can be calculated by simulation, or using the formula. To simulate, randomly sample n times from a binomial distribution with a given p, calculate $\hat{p}$, and repeat many times (e.g. 1000).

Ch. 6: tests for proportions that deal with a single category of a categorical variable. for categorical data, the parameter of interest is typically the population proportion. Ch. 7: tests involving two or more categories using χ^2 tests. frequency counts for the different categories of one or more categorical variables.

8.2 Standard error

Moral of the story: With enough information, simulation is not needed. Statistical theory provides the *actual* SE.

Sampling distribution of p is approximately normal if

- $0.25 \leq p \leq < 0.75$ (at 0, or 1, the tail is truncated)
- the sample size is large enough to keep the tail bound between 0 and 1, generally if $np \geq 10$ and $n(1-p) \geq 10$

The mean of the $\hat{p}$ is equal to the p, the probability of success.

$\hat{p} \sim N\left(p, sqrt(\frac{p(1-p)}{n})\right)$ if

- $0.25 \leq p \leq < 0.75$
- $np \geq 10$
- $n(1-p) \geq 10$

Therefore, if we only take a *single* sample, it is shown that given these conditions, this single sample comes from a distribution defined by $\hat{p} \sim N\left(p, sqrt(\frac{p(1-p)}{n})\right)$. This is useful, because with this one single sample, we can calculate SE directly as $\sqrt{\frac{p(1-p)}{n}}$, and substitute

If the conditions above do not hold, then a randomization test must be used.

Randomization distributions

Generate randomization samples that are consistent with the null hypothesis

When to use for the following?

- mean
- difference in means
- difference in proportions
- etc.

References