Week 3 Review

The 142 GSI Team

Introduction to Probability

- Look for patterns in language. Some words to look for are “given”, “and”, “or”, “both”, “neither”.
- Can you name more? Read some of your homework problems and see which words relate to which probability.
- Random samples eliminate sampling bias, but they can still provide an answer that is incorrect.
  - This is because when things are chosen at random, there is variation in the sample obtained.
  - If the variation when we take repeated samples from the same population is too great (really big), we can’t trust the results of any one sample

General Probability

- The probability of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions
    - We call a phenomenon random if individual outcomes are uncertain, but there is nonetheless a regular distribution of outcomes in a large number of repetitions.

Sample Space

- A sample space S of a random phenomenon is a set of all possible outcomes.
- An event on sample space S is an outcome (or a set of outcomes) of a random phenomenon. That is, an event is a subset of the sample space.

Example of discrete sample space: two outcomes (disease, not having the disease)

Example of continuous sample space: the different GPAs you can obtain

Probability Model

- A probability model is a mathematical description of a random phenomenon consisting of two parts: a sample space S and a way of assigning probabilities to events.

Probability Rules

- **Rule 1:** Probabilities are numbers between 0 and 1:

  $$\mathbf{0 \le P(A) \le 1}$$
- **Rule 2:** All possible outcomes of a sample space (S) together have a probability of 1:

  $$\mathbf{P(S) = 1}$$
- **Rule 3:** If two events have a joint probability of 0 (i.e., no overlap in their event spaces) then they are disjoint and the probability of either event occurring is the summation of their individual probabilities. If A and B are disjoint: P(A AND B) = 0

  $$\mathbf{P(AUB) = P(A) + P(B)}$$
  - This is the addition rule for disjoint events

- **Rule 4:** The probability of an event not occurring is 1 minus the probability of the event occurring. This event is called the complement and is denoted by a superscripted c or a single quotation mark.

  $$\mathbf{P(A') = P(A^c) = P(A\ does\ not\ occur) = 1 - P(A)}$$

Dependence and Independence

- Independent events
  - Two events A and B are independent if knowing that one event occurred does not change the probability that the other occurred
  - Multiplication Rule for Independent Events:
    - If A and B are independent, $P(A \cap B) = P(A) \times P(B)$
  - Conversely, if this condition is not satisfied, then events A and B are dependent.

- Dependent events
  - Two events A and B are dependent if knowing that one event occurred ***does*** change the probability that the other occurred
  - Multiplication Rule for Dependent Events:
    - If A and B are dependent, $P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A)$
- Disjoint events
  - Two events are disjoint if they c**annot both occur at the same time**.
  - This means the events are mutually exclusive.
  - If A and B are disjoint, $P(A \cap B) = 0$

More on Independent Events

- Independent events cannot be disjoint
- On a Venn diagram, independent events must overlap
- If two events overlap, they may or may not be independent
- To check if two events are independent you can perform any of the following checks:
    - Does P(A and B) = P(A) * P(B)? If it does, A and B are independent
    - Does P(A|B)=P(A)? If it does, then A and B are independent.
    - Does P(B|A)=P(B)? If it does, then A and B are independent.

More on Disjoint Events

- Two events are disjoint if they cannot both occur at the same time. This means the events are mutually exclusive
- Examples:
  - You get a test for COVID-19, you either test positive or test negative. You cannot test both positive or negative at the same time.
  - At 8:30am, you are either watching the lecture or sleeping (or something else!), you cannot both watch the lecture and be sleeping at the exact moment in time
- **Disjoint events are always dependent!** This is because knowing that the one event occurs implies there is a 0% chance the other event will also occur. This means P(A intersects B) = 0, or $P(A \cap B)=0$
- Using notation, P(test positive for COVID-19 | test negative for COVID-19) = 0%
- This proves the events are dependent because for independent events it must be the case that $P(A|B) = P(A)$. This says conditional on testing negative for COVID- 19, the probability that you test positive for COVID-19 is equal to the overall risk that you test positive for COVID-19. But we know this cannot be true because if you test negative then you cannot test positive on the same test!

Probability Calculation Rules for Any Two Events

- Conditional Probability
  - When $P(A) > 0$, the conditional probability of B given A is:

$$P(B|A) = P(A \,\&\, B) / P(A)$$

- General addition rule
  - For any two events A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

  - This simplifies to:

$$P(A \cup B) = P(A) + P(B) \text{ when A and B are disjoint}$$

- General multiplication rule
  - This is simply a rearrangement of the conditional probability formula:

$$P(A \cap B) = P(A)P(B|A)$$

  - This simplifies to:

$$P(A \cap B) = P(A) \times P(B) \text{ when A and B are independent}$$

- Bayes' Theorem
  - $P(A|B) = P(A)$ if independent. But if they aren't:

$$P(A|B) = P(B|A)P(A) / [P(B|A)P(A) + P(B|A^c)P(A^c)]$$

Probability Example: Hearing Impairment in Dalmations

Congenital sensorineural deafness is the most common form of deafness in dogs and is often associated with congenital pigmentation deficiencies. A study of hearing impairment in dogs examined over 5000 dalmatians for both hearing impairment and iris color. Being impaired was defined as deafness in either one or both ears. Dogs with one or both irises blue (a trait due to low iris pigmentation) were labeled blue. The study found that 28% of the dalmatians were hearing impaired, 11% were blue eyed, and 5% were hearing impaired and blue eyed.

Questions:
1. Use probability notation to write the known probabilities from the prompt.
2. What is the probability of dalmatians being either blue eyed or hearing impaired?
3. What is the probability of a dalmatian neither being impaired nor a blue-eyed dog?
4. Are blue eyes and impaired hearing independent or dependent? Why?

Solution

1. $P(I) = 0.28$
   $P(B) = 0.11$
   $P(B \cap I) = 0.05$

2. $P(B \cup I) = P(B) + P(I) - P(B \cap I) = 0.11 + 0.28 - 0.05 = 0.34$

3. $P(B^c \cap I^c) = P((B \cup I)^c) = 1 - P(B \cup I) = 1 - 0.34 = 0.66$

4. $P(B) * P(I) = 0.28 \times 0.11 = 0.0308$ $!= P(B \cap I) = 0.05$
   So the events B and I are dependent.

Diagnostic Testing

- **Sensitivity:** The test's ability to appropriately give a positive result when a person tested has the disease, or P(Test = + |Disease = TRUE)
    - [Big Idea: How good is our test at correctly identifying the disease-positive people?]
    - Try to think about this in terms of false positives/negatives
- **Specificity:** The test's ability to appropriately give a negative result when a person tested does not have the disease, or P(Test = - |Disease = FALSE)
    - [Big Idea: How good is our test at correctly identifying the disease-negative people?]
    - Try to think about this in terms of false positives/negatives
- **Positive predictive value:** The chance that a person truly has the disease, given that the test is positive, or P(Disease = TRUE|Test = +)
- **Negative predictive value:** The chance that a person truly does not have the disease, given that the test is negative, or P(Disease = FALSE |Test = -)

Example

- **Suppose that ABC Stats Inc. designed a COVID test that ALWAYS gives a positive result regardless of the disease status. (Hypothetically of course… this test would be useless in real life.) In reality, 2% of the population who take the test actually have COVID.**

Questions:

1. Are the results of the test and the disease status independent events?
2. Consider the three metrics for this test: sensitivity, specificity, and positive predictive value. Which of these values would be guaranteed to be 1? Which of these would be guaranteed to be 0? Calculate the remaining metric.
3. How would the values of sensitivity and specificity from question 2 change if the COVID test ALWAYS gave a negative result?

Solution

1. Yes, they are independent since knowing the result of the test does NOT change the probability of having the disease. (Or equally: knowing the disease status of the individual does not change the probability of getting a positive/negative test.)
2. In this case, since we have independence between the disease status and test results, we know that P(A|B) = P(A). [By the rules of independence]. Thus
   a. Sensitivity = P(Test = + |Disease = TRUE) = P(Test = +) = 1
   b. Specificity = P(Test = - |Disease = FALSE) = P(Test = -) = 0
   c. Positive Predictive Value = P(Disease = TRUE | Test = +) = P(Disease = TRUE) = 0.02
3. The values of the specificity and sensitivity would switch. (This is an example why the sensitivity and specificity of a test are both important values to look at. Sometimes healthcare scientists define a test based on some cutoff value of a body metric, like an individual's hormone levels, and picking the right cutoff value for the test involves balancing the test's sensitivity and specificity.)

Two by Two Tables: Sensitivity and Specificity

	Has Disease (D+)	No Disease (D-)	Total
Tests Positive (T+)	120	150	270
Tests Negative (T-)	30	750	780
Total	150	900	1050

Questions:

1. Use the table above to calculate the sensitivity, specificity, positive predictive value, and the negative predictive value.
2. Show how the sensitivity and positive predictive values are related to each other using Bayes' Theorem. (Could also do the same thing for negative predictive value and specificity)

Two by Two Tables: Sensitivity and Specificity

	Has Disease (D+)	No Disease (D-)	Total
Tests Positive (T+)	120	150	270
Tests Negative (T-)	30	750	780
Total	150	900	1050

Answers to Question 1

- Sensitivity = P(T+|D+) “Conditional on being diseased, what is the probability of testing positive?” 150 people have the disease, of which 120 test positive
  - Sensitivity = 120/150 = 4/5 = 80%
- Specificity = P(T-|D-) “Conditional on not being diseased, what is the probability of testing negative?” - 900 people do not have the disease, of which 750 test negative
  - Specificity = 750/900 =5/6=83.33%
- Patients want to know P(D+|T+) [positive predictive value] = “Conditional on the test being positive, what is the probability of having the disease?” 270 people tested positive, 120 of which had the disease.
  - Positive Predictive Value = 120 / 270 = 44.4%
- P(D-|T-) [negative predictive value] = “Conditional on the test being negative, what is the probability of not having the disease?” 780 people tested negative, 750 of which had the disease
  - Negative Predictive Value = 750 / 780 = 96.2%

Answers to Question 2

Applying Baye's Theorem

PPV : $P(D+|T+) = \dfrac{P(T+|D+)\,P(D+)}{P(T+|D+)\,P(D+) + P(T+|D-)\,P(D-)}$ where $P(T+|D+)$ is the sensitivity

NPV : $P(D-|T-) = \dfrac{P(T-|D-)\,P(D-)}{P(T-|D-)\,P(D-) + P(T-|D+)\,P(D+)}$ where $P(T-|D-)$ is the specificity

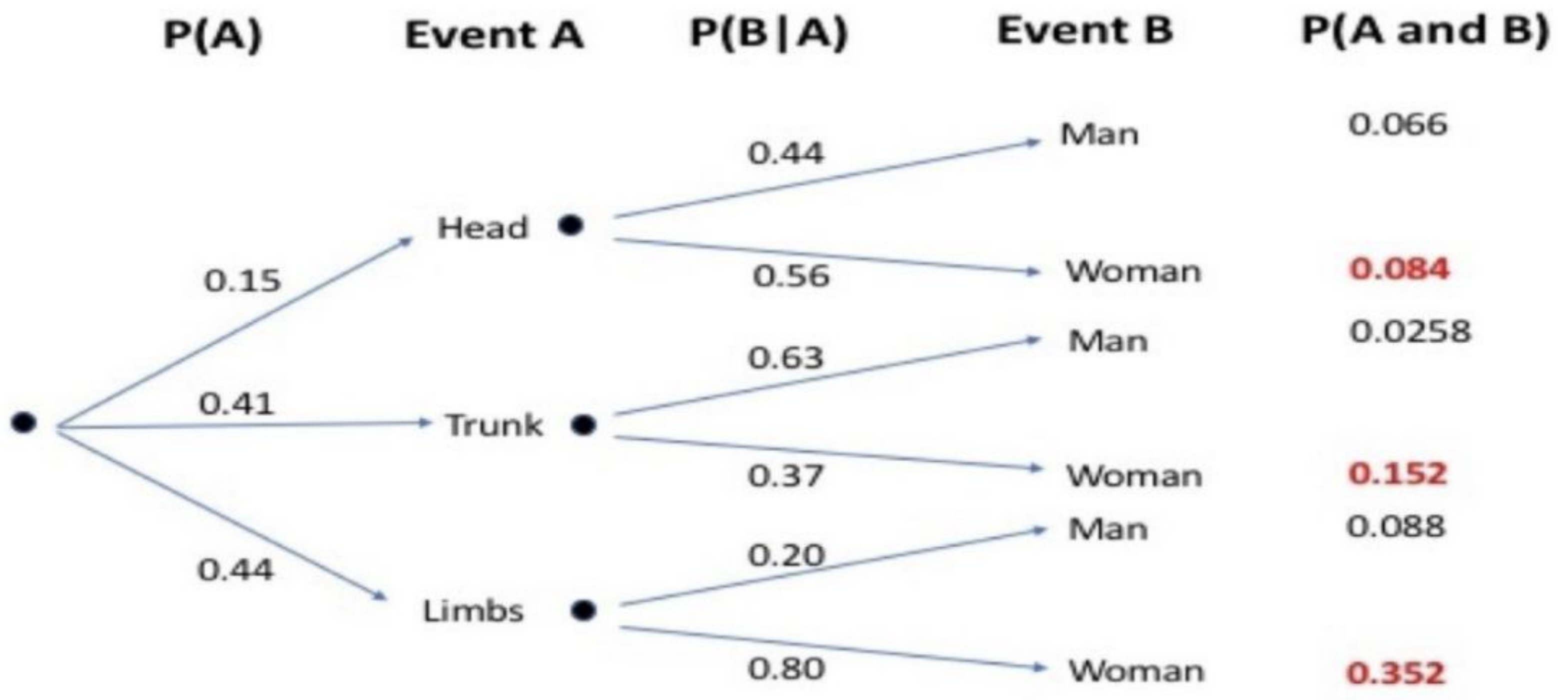
Example: Absolute Frequencies and Tree Diagrams

A study of the presence of cutaneous malignant melanoma at a single body location among the Italian population found that 15% of skin cancers are located on the head and neck area, another 41% on the trunk and the remaining 44% on the limbs. 44%of individuals with skin cancer on the head are men, as are 63% of those with skin cancer on the trunk but only 20% of those with skin cancer on the limbs. **What percent of all individuals with skin cancer are women?**

$P(\text{head}) = 0.15$
$P(\text{trunk}) = 0.41$
$P(\text{limbs}) = 0.44$

$P(\text{man}|\text{head}) = 0.44$
$P(\text{man} \mid \text{trunk}) = 0.63$
$P(\text{man} \mid \text{limbs}) = 0.20$

Solution: Tree Diagram



Solution: Probability Notation

P(man | head) = 44%
P(woman | head) = 100 - 44 = 56%
similarly, P(woman | trunk) = 37%
P(woman | limbs) = 80%

P(woman) = P(head)P(woman|head) + P(trunk)P(woman|trunk) + P(limbs)P(woman|limbs) = (15% * 56%) + (41% * 37%) + (44% * 80%) = 58.5%

Two by Two Tables: Risk and Odds

	Has Disease (D+)	No Disease (D-)	Total
Is Exposed (E+)	100	50	150
Not Exposed (E-)	50	300	350
Total	150	350	500

- What is the **risk (a.k.a. probability)** that someone has the disease?
  - Note, this questions asks about a marginal probability. P(Has disease) = P(D+) = 150/500=0.3=30%, call this probability "p".
  - Risk of having the disease in this population is 30%
- What is the **odds** that someone has the disease?
  - **Odds of having the disease:** p/(1-p) = (150/500)/(350/500) = 150/350 = 3/7
  - Alternate calculation: Odds of having the disease: p/(1-p) = 0.3/0.7= 3/7 "3 to 7 odds of having the disease)

Normal Distribution

Distributions

- Can think of distributions as things that model patterns in data
- We can describe our data using different distributions like Normal, Binomial, Poisson
    - Example: we can calculate the distribution’s mean, variance/standard deviation, and calculate probabilities based on what we know about the data’s underlying distribution

Normal Distribution

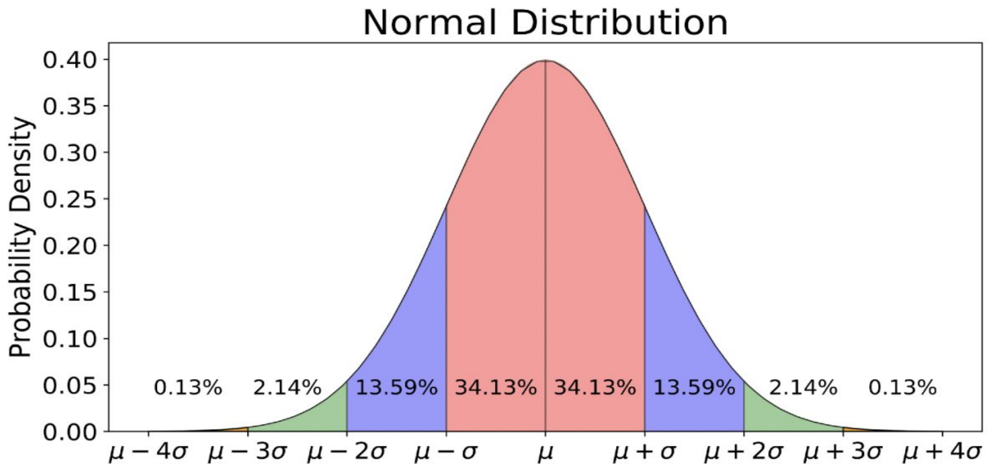


Figure 1: plot of chunk unnamed-chunk-6

Normal Distribution: A Continuous Distribution

1. Goals: We want to find a probability for greater than or less than or greater/equal, or less/equal,or falling within a certain interval, but NEVER exactly equal to a single value
2. The data are said to be normal or are assumed to be normal in the prompt

Normal Distribution Functions

Function	Use	Example
`pnorm()`	Compute the probability at a specified x or below. Is a cumulative probability. Need to also specify the mean and sd. Remember P(X=2) = 0! This holds for any value "x".	P(X<=2): `pnorm(2, mean = 3, sd = 0.5)` P(X<2): `pnorm(2, mean = 3, sd = 0.5)`
`rnorm()`	Generate a random variable that is drawn from a normal distribution. We use this is create/simulate data to get a sense of what Normally distributed data looks like	`rnorm(10, mean = 3, sd = 0.5)` generates 10 random draws from a Normal distribution
`qnorm()`	Is the opposite of pnorm(). You provide a percentage and it returns the x value such that x percent of the area is at/below that value in the lower tail.	P(X<=x) = 75%, what is x? `qnorm(0.75, mean = 3, sd = 0.5)` will return the value x.

Normal Distribution

- Remember for the Normal distribution: $P(X = x) = 0$
  - example: $P(X = 2) = 0$
  - Where x can be any value (here it is 2).
  - This property is true for any continuous distribution – the only one we study is the Normal distribution.
- This is NOT true for Binomial and Poisson; they can have non-zero probabilities of being exactly equal to whole numbers (like 2).
  - Thus, $P(X<=2) = P(X<2)$ for the Normal distribution.
  - For Binomial or Poisson, $P(X<2) = P(X = 0) + P(X = 1)$, but $P(X<=2) = P(X = 0) + P(X = 1) + P(X = 2)$

Questions

- What is the difference between normal and standard normal?
- what is a z-score? How do you interpret a z-score?
- what is the 68-95-99.7 rule for the normal distribution?

Example: Normal Distribution

For babies born at full term (37 to 39 completed weeks of gestation), the distribution of birth weight (in grams) is approximately normally distributed with a mean of 3350 grams and a standard deviation of 440 grams, N(3350,440).

1. **What is the probability that a baby is born with a weight less than 3790 grams?**
2. **What weight corresponds to the 68th percentile of baby birth weights?**
3. **Use code to generate 100 observations from a normal distribution where the mean is 3350 grams and sd = 440 grams.**

Solution

The prompt information translates to:

```
mu <- 3350
sigma <- 440
```

We can calculate this probability one way:

```
prob <- pnorm(3790, mean=mu, sd=sigma)
percentile <- qnorm(0.68, mean=mu, sd=sigma)
birth.weights <- rnorm(100, mean=mu, sd=sigma)
```

All the [ ]norm functions (except dnorm()) are fair game!

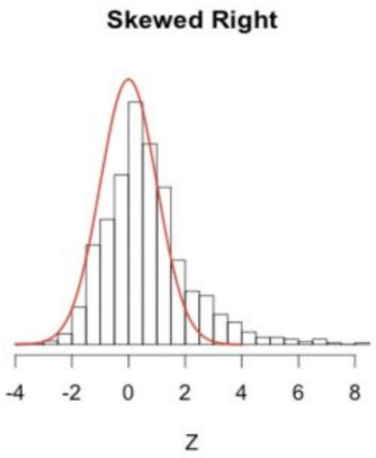
Solution

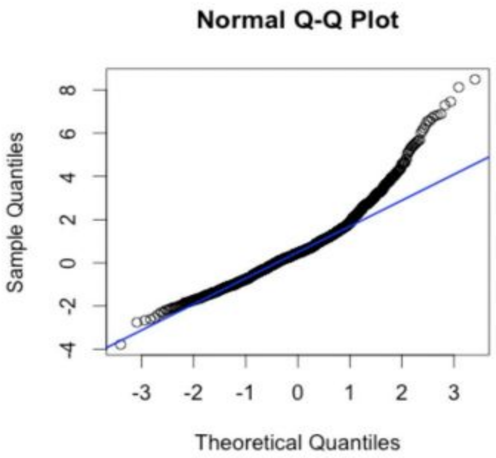
Use a histogram and a qq-plot to test whether your data is normal. In our `birth.weight` data, the variable is named `weight`.

```
plot.hist <- ggplot(data = birth.weights, aes(x=weight)) +
    geom_histogram()
plot.qq <- ggplot(data = birth.weights, aes(sample=weight)) +
    stat_qq() +
    stat_qq_line()
```

QQ Plots

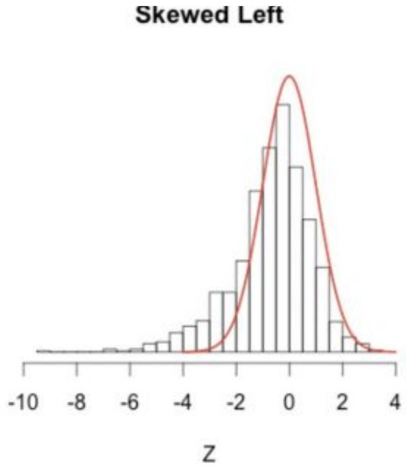
Skewed right data deviates positively in the right top corner. If data were normally distributed, 99.7% would have sample quantiles between -3 and 3, but here the sample quantiles are between 4 and 8 -> way outside of the range for normal data for the data on the right-hand side of the histogram.

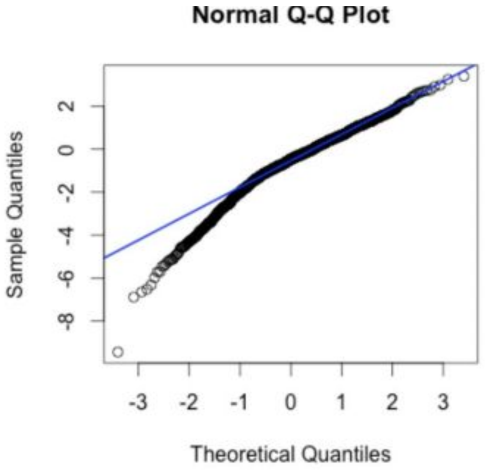




QQ Plots

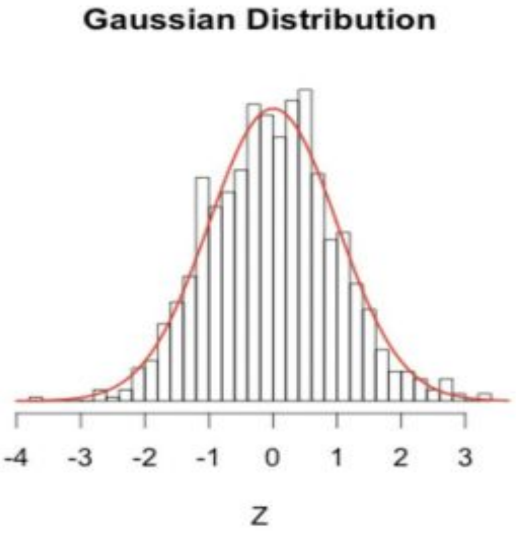
Skewed left data deviates negatively in the left bottom corner. If data were normally distributed 99.7% would have sample quantiles between -3 and 3, but here the sample quantiles are between -4 and -8 way outside of the range for normal data for the data on the left hand side of the histogram

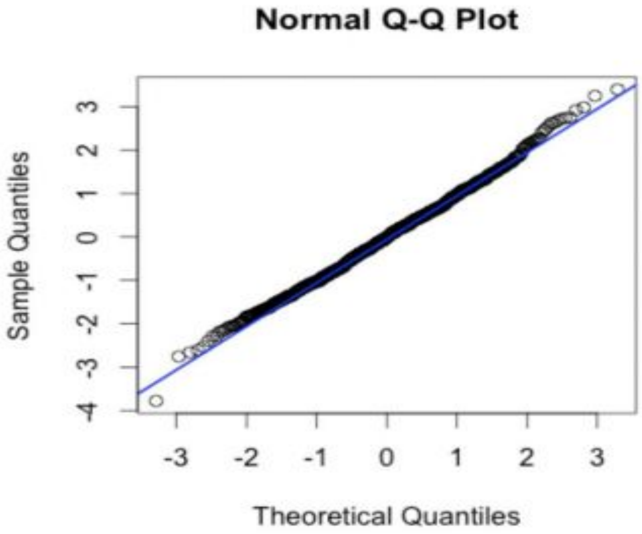




QQ Plots

Normal data falls on straight line in a QQ plot





Binomial Distribution

Binomial Distribution

- Distribution for discrete random variables (those that can only take whole numbers. This is different from Normal, which is for continuous random variables)
- To determine whether something is binomial, look for the “n” and “p” and whether the trials are independent.
  - n: number of “trials”
  - p: probability of success
- Normal approximation to the binomial: When n is large the binomial distribution will look like a normal distribution.
  - Use when $n*p >= 10$ and $n*(1-p) >= 10$

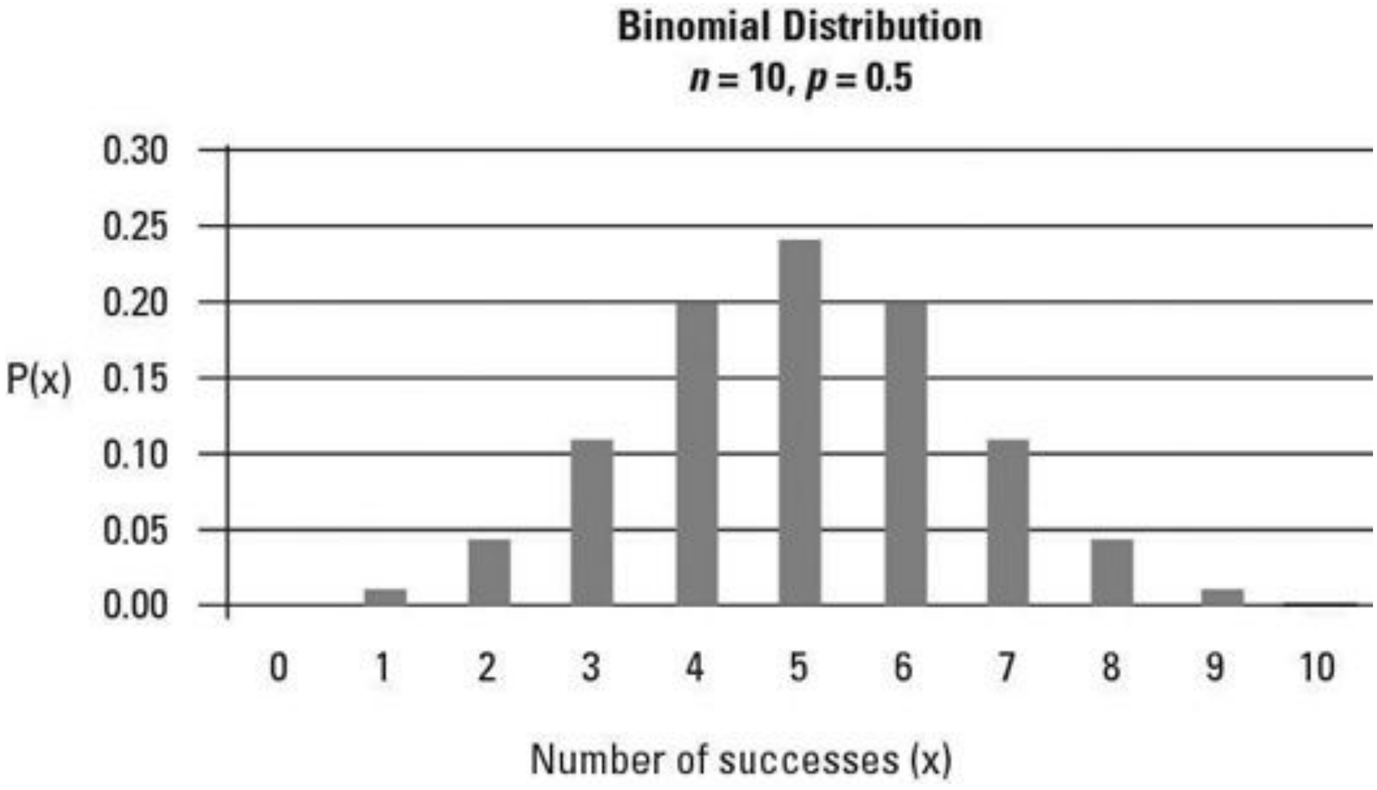
Binomial Distribution

This is the formula for the Binomial distribution:

$$P(X = k) = \binom{n}{k} * p^k * (1-p)^{n-k}$$

- Understand what each piece of the binomial distribution function is
- As we said earlier, distributions are used to calculate means, variances, and probabilities of situations we see often! Think back to the Korean drama or pop song. They have well-known structures. So does the Binomial setting. Check if your data fit the Binomial setting.

Binomial Distribution



Binomial Functions

Function	**Use**	**Example**
`pbinom()`	calculating the cumulative probabilities	To compute P(X<=2), where X follows a Binomial(10, 0.5) we use pnorm(2, 10, 0.5)
`dbinom()`	calculating the probability mass function	To compute the mass of a Binomial(10, 0.5) at 2, we use dnorm(2, 10, 0.5)
`rbinom()`	generate random samples from a binomial distribution	To generate 3 random numbers from a Binomial(10, 0.5), we use rbinom(3, 10, 0.5)
`qbinom()`	We didn’t discuss this and you don’t need to know it	/

Binomial Distribution

Goals: We want to find a probability:

- less than
- equal
- greater than
- combo

Must meet these assumptions:

1. You have some fixed amount of trials
2. Trials are independent
3. The probability of success is the same for each trial
4. There are only two outcomes: success or failure

Normal Approximation for the Binomial Distribution

- If we have a binomial distribution that satisfies the below properties, we can approximate the distribution as a normal distribution
- This allows us to use all the awesome properties we know about normal distributions (very convenient!)

NORMAL APPROXIMATION FOR BINOMIAL DISTRIBUTIONS

Suppose that a count X has the binomial distribution with n observations and success probability p. When n is large, the distribution of X is approximately Normal, $N(np, \sqrt{np(1-p)})$.

As a rule of thumb, we will use the Normal approximation when n is so large that $np \geq 10$ and $n(1-p) \geq 10$.

Example

The probability that a patient recovers from a widespread flu is 90%. There are now 30 people in a hospital who have contracted the flu independently. What is the probability that 25 patients in that hospital will recover from it? How about more than 25 patients?

Solution

The prompt information translates to:

```
n_trials <- 30
k_success <- 25
probability <- 0.9
```

And we can calculate the first quantity “by hand” using the formula:

```
choose(n_trials, k_success)*(probability)^k_success*(1-probability)^(n_trials-k_success)
```

[1] 0.1023048

Or we can use the `dbinom()` function in R:

```
dbinom(x=k_success, size=n_trials, prob=probability)
```

Solution (cont.)

The second quantity (the probability of more than 25 patients) can be found using the formula for the binomial distribution. Note the difference between “more than” (AKA >25) and “no less than” (AKA >=25)

```
choose(n_trials, 26)*(probability)^(26)*(1-probability)^(n_trials-26) +
choose(n_trials, 27)*(probability)^(27)*(1-probability)^(n_trials-27) +
choose(n_trials, 28)*(probability)^(28)*(1-probability)^(n_trials-28) +
choose(n_trials, 29)*(probability)^(29)*(1-probability)^(n_trials-29) +
choose(n_trials, 30)*(probability)^(30)*(1-probability)^(n_trials-30)
```
[1] 0.8245051

Or using another cool function that calculate the sum of probabilities from 0 to k and take the complement.

```
1-pbinom(q=k_success, size=n_trials, prob=probability)
```
[1] 0.8245051

Or even more conveniently:

```
pbinom(q=k_success, size=n_trials, prob=probability, lower.tail=FALSE)
```
[1] 0.8245051

Binomial Distribution Key Takeaways

- Know the formula to calculated probabilities when a random variable follows a binomial distribution
- Know when you can apply the Normal approximation and how to do it
- Know the R code to calculate binomial probabilities

Poisson Distribution

Poisson Distribution

This is the formula for the Poisson distribution with mean (or rate) lambda.

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Goals: We want to find a probability:

- greater
- less
- equal
- combo

Must meet these assumptions:

1. Events occur independently.
2. The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals.

Poisson Functions

Function	Use	Example
`ppois()`	Compute the probability at a specified x or below (lower.tail=T by default). Is a cumulative probability. Need to also specify lambda, the mean of the Poisson distribution	P(X<=2): ppois(2, lambda = 0.5) P(X<5): ppois(4, lambda = 0.5) P(X>5) = 1-P(X<=5) = 1 – ppois(5, lambda = 0.5)
`dpois()`	Compute an exact probability at a specified value for x, where x is a discrete value between 0 and infinity. Need to also specify lambda, the mean of the Poisson distribution	P(X=2): dpois(2, lambda = 0.5) P(X=1) + P(X=2) (can add them together)
`rpois()`	Generate a random variable that is drawn from a Poisson distribution. We use this is create/simulate data to get a sense of what Poisson data looks like	rpois(10, lambda = 0.5) generates 10 random draws from a Poisson distribution
`qpois()`	We didn’t discuss this and you don’t need to know it	\

Example

Suppose there is a disease that occurs at a rate of 10 per million people for a period of one year. Let X be the number of cases in 1 million people each year. What is the probability that we see more than 12 cases of the disease in a city within one year? How about exactly 12?

Solution

What is the probability that we see more than 12 cases of the disease in a city of 1 million people within one year?

```
> ppois(q=12,lambda=10,lower.tail=F)
[1] 0.2084435
```

How about exactly 12? Practice calculating this using the probability distribution function by hand as well.

```
> dpois(x=12, lambda=10)
[1] 0.09478033
```

Calculating upper tails using Poisson/Binomial

Suppose you want to calculate P(X>3). How would you do it?

1. Calculate 1 – P(X<=3) which is coded by:

```
1-ppois(3, lambda=0.5)
```

2. Alternatively, use lower.tail = F. But be careful! Does ppois(3, lambda = 0.5, lower.tail=F) give you P(X>=3) or P(X>3)?

   It gives you P(X>3). Can check this by ensuring that ppois(3, 0.5) + ppois(3, 0.5, lower.tail = F) = 1, which means these two probabilities do not overlap. The first term includes 3 in the probability calculation but the second one does not. The same thing holds for Binomial calculations.