Лекция 3

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Базовая описательная статистика

Здравствуйте!

После изучения материала лекции, вы сможете:

1. Объяснить, как описательная статистика может сделать данные понятными.
2. Строить и анализировать плотности распределения для переменных в каждом из трех уровни измерения.
3. Вычислять и интерпретировать проценты, пропорции, отношения, ставки и процентное изменение.
4. Анализировать графики.

Результаты исследований не говорят сами за себя. Исследователи используют статистику для организации и манипулирования данными, чтобы их читатели могли понять их значение. Целью описательной статистики является четкое и эффективное выражение результатов исследований. Сейчас мы рассмотрим некоторые часто используемые методы представления результатов исследований, включая таблицы, проценты, показатели и графики. Эти одномерные описательные статистические данные не являются математически сложными (хотя они не так просты, как могут показаться на первый взгляд), но они могут быть чрезвычайно полезными инструментами для организации и анализа результатов и передачи выводов.

Очень часто первым шагом в проекте количественного исследования является изучение переменных и просмотр распределения баллов. Один из наиболее полезных способов сделать это – построить таблицы или распределения частот, которые сообщают о количестве. Что же такое частотное распределение?

Частотное распределение, это метод статистического описания данных (измеренных значений, характерных значений). В описательной статистике частота распределения имеет ряд математических функций, которые используются для выравнивания и анализа частотного распределения.
Построение распределений для переменных номинального уровня не очень сложный процесс. Подсчитайте, сколько раз встречается каждая категория или оценка переменной, и отобразите частоты в табличном формате. Например, в таблице показано распределение пола по 113 респондентам.

Таблица 3.1 Пол (фиктивные данные)

Пол	Количество
Женщины	60
Мужчины	53
Итого	113

Обратите внимание, что в таблице есть заголовок и пометка, и в нижней части столбца частоты указывается общее количество случаев. Эти элементы должны быть включены во все частотные распределения.

Таблицы 3.2 и 3.3 иллюстрируют выбор, который когда-либо должен быть сделан. Таблица 3.2 отображены «стандартные» пять основных направлений в образовании Республики Казахстан на сегодня. В Таблице 3.3 представлены более подробные данные, добавленные к трем крупнейшим

«Другим». Число категорий будет еще больше, если мы включим больше «других» видов образований.

В этот момент у нас возникнет вопрос. Достаточно ли детализирована информация? Когда картина, представленная в таблице, становится слишком загроможденной, слишком сложной и неясной? Это вопросы, на которые необходимо ответить в контексте цели исследовательского проекта. Если вы хотите подчеркнуть численное преобладание людей с высшим образованием в Республике Казахстан, предпочтение следует отдать Таблице 3.2.

Таблица 3.2 Количество человек, получающих образование в Республике Казахстан, 2018 г.

Образование	Количество (чел)
Дошкольное образование	862 305
Среднее образование	3 050 800
Профессиональное образование	64 814
Высшее образование	542 458
Другое	537 917
Итого	5 058 294

Источник – Комитет по статистике Республики Казахстан http://stat.gov.kz/faces/wcnav_externalId/homeNumbersEducation?_afrLoop=5688560415843310#%40%3F_afrLoop%3D5688560415843310%26_adf.ctrl-state%3Ddbysid1pf_145

Таблица 3.3 Количество человек, получающих образование в Республике Казахстан, 2018 г.

Категория образования	Количество (чел)
Дошкольное образование	862 305
Среднее образование (дневное)	3 039 100
Среднее образование (вечернее)	75 000
Профессиональное образование	64 814
Колледж	489 337
Высшее образование	542 458
Послевузовское образование	48 580
Итого	5 121 594

Источник – Комитет по статистике Республики Казахстан http://stat.gov.kz/faces/wcnav_externalId/homeNumbersEducation?_afrLoop=5688560415843310#%40%3F_afrLoop%3D5688560415843310%26_adf.ctrl-state%3Ddbysid1pf_145

С другой стороны, если вы хотите подчеркнуть разнообразие образования, таблица 3.3 будет предпочтительнее. Не существует жестких правил, и выбор между большей детализацией (больше категорий) и большей ясностью (меньше категорий) может столкнуться с переменными на всех трех уровнях измерения.

Вероятно, вы заметили трудность работы с таблицами 3.2 и 3.3: используемые частоты велики и, следовательно, могут быть трудны для понимания. Например, представьте, что вы пытаетесь объяснить эти таблицы в терминах частот: «3 039 100» из 16 367 000 имеют среднее образование. «В этом утверждении нет ничего плохого - оно в буквальном смысле верно, но то же самое можно было более четко выражено в процентах: «Около 54% казахстанцев имеют среднее образование». Проценты являются чрезвычайно полезной статистикой, потому что они обеспечивают систему отсчета, стандартизируя необработанные частоты для базы 100. Математическое определение процента:

Формула 3.1 Процент: $\%=(fN)\times 100$

Где f = частота или число случаев в определенной категории,
N = число случаев во всех категориях.
Для иллюстрации расчета процентов рассмотрим данные, представленные в таблице 3.1. Как мы можем найти процент мужчин в выборке? Обратите внимание, что 53 мужчины (f= 53) и 113 случаев (N = 113). Так

$$\%=fN\times 100=53113\times 100=0.4690\times 100=46.9\%$$

Используя ту же процедуру, мы можем найти процент женщин:

$$\%=fN\times 100=60113\times 100=0.5310\times 100=53.10\%$$

Проценты воспринимать легче, чем необработанные частоты, и столбец для процентов обычно добавляется к частотным распределениям для переменных на всех уровнях измерения. Например, в Таблице 3.4 представлена та же информация, что и в Таблице 3.3, но проценты значительно облегчают чтение.

Таблица 3.4 Количество человек в (%) получающих образование в Республике Казахстан, 2018 г.

Категория образования	Количество	в %
Дошкольное образование	862 305	16,8
Среднее образование (дневное)	3 039 100	59,3
Среднее образование (вечернее)	75 000	1,5
Профессиональное образование	64 814	1,3
Колледж	489 337	9,6
Высшее образование	542 458	10,6
Послевузовское образование	48 580	0,9
Итого	5 121 594	100,0

Источник – Комитет по статистике Республики Казахстан http://stat.gov.kz/faces/wcnav_externalId/homeNumbersEducation?_afrLoop=5688560415843310#%40%3F_afrLoop%3D5688560415843310%26_adf.ctrl-state%3Ddbysid1pf_145

Преимущество использования процентов особенно очевидно, когда мы хотим сравнить группы разных размеров. Например, рассмотрим информацию, представленную в таблице 3.5. В каком колледже больше специальностей по общественным наукам? Поскольку общее количество специальностей очень разное, сравнивать их по необработанным частотам сложно.

Таблица 3.5 Количество студентов по специальностям в двух студенческих городках (фиктивные данные)

Специальности	Колледж А	Колледж В
Бизнес	103	3120
Естественные науки	82	2799
Общественные науки	137	1884
Гуманитарные науки	93	2176
Итого	415	9979

Таблица 3.6 Количество студентов в (%) по специальностям в двух колледжах (фиктивные данные)

Специальности	Колледж А	Колледж В
Бизнес	24,8	31,3
Естественные науки	19,8	28,0
Общественные науки	33,0	18,9
Гуманитарные науки	22,4	21,8
Итого	100,0	100,0

Таблица 3.6 упрощает выявление как различий, так и сходств. Так в колледже А имеет более высокий процент специальностей по общественным наукам, чем колледж В (хотя абсолютное число специальностей по общественным наукам меньше, чем в колледже В) и примерно такой же процент гуманитарных специальностей. Как бы вы описали различия в оставшихся двух основных областях?

Социологи используют отношения, пропорции и проценты. Пропорции варьируются от 0,00 до 1,00: они стандартизируют результаты на основе 1,00 вместо – на основе 100, используемой для процентов. Пропорция равна проценту, за исключением того, что мы не умножаем на 100.

Формула 3.2 Пропорции: =fN

Проценты могут быть преобразованы в пропорции путем деления на 100, и, наоборот, пропорции могут быть преобразованы в проценты путем умножения на 100. Эти две статистики являются эквивалентными выражениями одного и того же вычисления и могут взаимозаменяться. Например, мы могли бы указать относительное количество мужчин в Таблице 3.1 как пропорцию:

Пропорции: =fN=53113=0,47

Как мы можем выбирать между показателями? Проценты легче понять большинству людей (включая статистиков) и, как правило, предпочтительнее. Пропорции используются реже, как правило, когда мы работаем с вероятностями. Предпочтение процентам основано исключительно на простоте общения: пропорции и проценты являются одинаково допустимыми способами выражения результатов. Рассмотрим распределение частот для переменных порядкового уровня.

Распределения частот для переменных порядкового уровня строятся так же, как и для переменных номинального уровня. В таблице 3.7 представлены данные о распределении частот ответов учащихся в проведенном исследовании, в котором измеряется возможность введения процедуры рождаемости в университетском городке. Обратите внимание, что процентный столбец был добавлен в таблицу для повышения ясности.

Таблица 3.7 Контроля над рождаемостью в университетском городке (фиктивные данные)

Ответ	Количество	Процент
Сильно согласны	350	25.55%
Согласны	462	33.72%
Не согласны	348	25.40%
Категорически не согласны	210	15.33%
Итого	1370	100%

(обеспечение презервативами)

Таблица 3.8 Контроля над рождаемостью в университетском городке (фиктивные данные)

Ответ	Количество	Процент
Сильно согласны и Согласны	812	59.27%
Не согласны и Категорически не согласны	558	40.73%
Итого	1370	100%

Как видно из таблицы 3.8 мнения были распределены достаточно равномерно. Самым популярным ответом было «Согласен» (33,72%), и большинство студентов (59,27%) согласились или полностью согласились с тем, что презервативы и другие устройства «безопасного секса» должны быть доступны. Если исследователь хочет подчеркнуть эту закономерность или сделать таблицу более компактной, категории можно свернуть, как показано в таблице 3.8. Однако за более простое и более компактное выражение результатов приходится платить: точная разбивка степеней согласия и несогласия теряется. Как мы видели при обсуждении образования в таблицах 3.2 и 3.3 исследователь должен найти баланс между большей детализацией (больше категорий) и большей ясностью (меньше категорий). Рассмотрим теперь распределение частот для интервала переменного уровня. В общем, построение частотных распределений для переменных, измеренных на уровне отношения интервалов, является более сложным, чем для номинальных и порядковых переменных. Переменные с интервалами обычно имеют широкий диапазон баллов, и это означает, что исследователь должен свернуть или сгруппировать категории, чтобы получить достаточно компактные таблицы. Еще раз, мы видим, что исследователь должен сделать вывод между большей детализацией и большей ясностью.

Предположим, что вы хотите сообщить о распределении переменной «возраст» для сообщества. В большинстве сообществ будет очень широкий диапазон возрастов, от новичков до людей в возрасте от 90 лет и старше. Если вы просто сообщаете о количестве единиц, которые произошли в каждом году (или балле), вы можете получить распределение частот по категориям 80, 90 или даже несколько раз; такую таблицу было бы очень трудно читать. Баллы (годы) должны быть сгруппированы в большие категории, чтобы увеличить доступность? Насколько большими должны быть эти категории? Сколько категорий должно быть включено в таблицу? Должны ли мы предоставить больше информации (большее количество узких категорий) или большую ясность (меньшее количество широких категорий)? Данные вопросы остаются открытыми. Имеется также возможность построения частотных распределений для переменных уровня отношения. Для простоты рассмотрим распределение частот для небольшой группы из 20 студентов. Из-за более узкого возраста студентов, мы можем использовать категории только одного года (эти категории часто называют интервалами классов при работе с данными отношения интервалов). Распределение частот строится путем упорядочения возрастов, подсчета количества раз, когда встречается каждый счет (год возраста), и затем суммирования количества случаев в каждой категории. Таблица 3.9 представляет информацию и показывает концентрацию случаев в 18 и 19 классовых интервалах. Несмотря на то, что эта таблица довольно ясна, предположим, что вы хотите более компактное (менее подробное) резюме.

Таблица 3.9 Возраст Студентов в группе колледжа (фиктивные данные)

Возраст	Количество
18	5
19	6
20	3
21	2
22	1
23	1
24	1
25	0
26	1
Итого	20

Чтобы получить это, вам нужно сгруппировать результаты в более широкие интервалы между опрошенными. Увеличение ширины интервала (скажем, до двух лет) уменьшит количество интервалов и даст более компактное выражение. Группировка баллов в таблице 3.9 четко подчеркивает преобладание. Эту тенденцию можно подчеркнуть еще больше, добавив столбец для процентов.

Таблица 3.10 Возраст Студентов в Группе Колледжа (фиктивные данные)

Возраст	Количество	%
18-19	11	55%
20-21	5	25%
22-23	2	10%
24-25	1	5%
26-27	1	5%
Итого		20

Обратите внимание, что интервалы в Таблице 3.10 указаны с очевидным разрывом между ними. То есть заявленные пределы разделены расстоянием в одну единицу и это допустимо.

Рассмотрим теперь понятия кумулятивная частота и кумулятивный процент. Эти два понятия обычно используются как дополнения к основному распределению частот для данных между коэффициентами являются столбцы кумулятивной частоты и кумулятивного процента. Что же это такое? Накопленная частота – это сумма частот данного и всех предшествующих интервалов. Куммулята позволяет определить, какая часть совокупности обладает значениями изучаемого признака не превышающими заданного предела, а какая часть превышает этот предел. Эти добавленные в таблицу столбцы позволяют с первого взгляда определить, сколько случаев попадает в распределение или ниже заданного значения или интервала между классами.

Таблица 3.11 Возраст студентов в классе колледжа

Возраст	Количество	Кумулятивная частота	Процент	Кумулятивный процент
18-19	11	11	55.0%	55.0%
20-21	5	16	25.0%	80.0%
22 — 23	2	18	10.0%	90.0%
24-25	1	19	5.0%	95.0%
26-27	1	20	5.0%	100.0%
Итого	20		100.0%	

Кроме того, эти накопительные столбцы весьма полезны в ситуациях, когда исследователь хочет высказать мнение о том, как случаи распределяются по диапазону оценок. Например, таблица 3.11 ясно показывает, что подавляющее большинство учащихся в классе моложе 21 года. Если исследователь хочет подчеркнуть этот факт, то эти накопительные столбцы весьма удобны.

Наиболее реалистичные исследовательские ситуации будут касаться более чем 20 случаев и/или гораздо большего числа катетеризаций, чем в наших таблицах, и столбец совокупного процента обычно предпочтительнее столбца совокупных частот.

Как правило, интервалы классов частотных распределений должны быть одинаковыми, чтобы максимизировать ясность и простоту понимания. Например, все интервалы класса в таблице 3.11

имеют ширину 2 года. Есть несколько других возможностей для определения интервалов между классами, и мы рассмотрим каждую ситуацию отдельно.

Что произойдет с распределением частот в «3.11», если мы добавим одного ученика, которому было 47 лет? У нас будет 21 случай, и между самым старым респондентом (47 лет) и вторым самым старшим (26 лет) будет большой разрыв. Если бы мы просто добавили старшего ученика, нам пришлось бы включить девять новых интервалов (28–29, 30–31 и т.д.) С нулевыми значениями в них, прежде чем мы дойдем до интервала 46–47. Альтернативой для обработки нескольких очень высоких (или низких) баллов было бы добавление «открытого» интервала в соответствии с распределением частот, как в таблице 3.12.

Таблица 3.12 Возраст студентов в группе (классе) колледжа

Возраст	Количество	Совокупное количество
18-19	11	11
20-21	5	16
22 — 23	2	18
24-25	1	19
28 и более	1	21
Итого	21	

Открытый интервал в Таблице 3.12 позволяет нам представить информацию в достоверном виде. Мы могли бы справиться с чрезвычайно низкой оценкой, добавив открытый интервал между самым низким классовым интервалом (например, «17 и младше»). Конечно, эта эффективность имеет цену – в таблице не указана точная оценка.

Открытый интервал – поэтому этот метод не должен использоваться без детализации. Следующим вопросом нашей лекции являются понятия соотношения, ставки и процентное изменение. Соотношения, ставки и процентные изменения – это статистика, которая используется для простого и ясного подведения итогов. Они могут использоваться независимо или с частотным распределением, и они могут быть рассчитаны для переменных на любом уровне измерения. Хотя они похожи друг на друга, у каждой статистики есть конкретное применение и цель, и мы будем рассматривать их по одному.

Соотношения рассчитываются путем деления частоты в одной категории на частоту в другой. Формула для соотношения

Формула 3.3 Соотношения =f1f2

где f1 = количество дел в первой категории
f2 = количество дел во второй категории

Соотношения особенно полезны для сравнения относительных размеров различных категорий переменной. Чтобы проиллюстрировать это, предположим, что вас интересовали относительные соотношения мужчин и женщин. Чтобы найти отношение женщин (f1) к по отношению к (f2),

разделите 60 на 53:

Соотношение=f1f2=6053=1.13

Еще одно понятия ставки, предоставляют еще один способ суммирования распределения одной переменной. Скорость определяется как количество фактических случаев возникновения явления, деленное на количество возможных случаев за определенную единицу времени. Значения обычно умножаются на некоторую степень 10, чтобы исключить десятичные точки. Например, общий коэффициент смертности для населения определяется как общее количество смертей в этой группе (фактические случаи), деленное на количество людей в группе (возможные случаи) в год. Эта величина затем умножается на 1000. Эта формула может быть выражена как

Общий коэффициент смертности = (Количество смертей/Общую численность населения) *100

Что же такое процентное изменение?

Измерение социальной ценности во всем ее многообразии является важной задачей для всех общественных наук. Одна очень полезная статистика для этой цели - процентное изменение, которое говорит нам, насколько переменная увеличилась или уменьшилась за определенный промежуток времени.

Чтобы вычислить эту статистику, нам нужны оценки переменной в два разных момента времени. Баллы могут быть в форме частот, скоростей или процентов. Изменение в процентах скажет нам, насколько изменилась оценка в более позднее время по сравнению с более ранним временем. Используя в качестве примера показатели смертности, представьте, что общество страдает от разрушительной вспышки болезни, в которой уровень смертности вырос с 16 смертей на 1000 в 2000 году до 24 смертей на 1000 в 2010 году. Очевидно, что уровень смертности выше в 2010, но насколько относительно 2000?

Формула 3.4 для процентного изменения

Процентное изменение = (f1-f2) : f1×100

где f1 - первая оценка, частота или значение
f2- вторая оценка, частота или значение

В нашем примере f1 – уровень смертности в 2000 году (f1 =16) и f2 – уровень смертности в 2010 году (f2 =24). Формула говорит нам вычесть более раннюю оценку из более поздней оценки, а затем разделить на более раннюю оценку. Результат выражает размер изменения в баллах (f2 - f1) относительно балла в более раннее время (f1). Затем значение умножается на 100, чтобы выразить изменение в виде процента.

Процентное изменение = (f1-f2):f1 = (24-26):16×100 = (18:16)×100 = 0.50 × 100 =50%

Не менее интересным вопросом является использование графиков для представления данных.

Исследователи часто используют графики для визуального представления своих данных. Эти инструмент особенно полезны для сообщения общей формы распределения и для выделения любой кластеризации случаев в определенном диапазоне баллов.

Доступно много типов графиков, но мы рассмотрим только два. круговые диаграммы и

гистограммы, подходят для переменных на любом уровне измерения. Все описанные здесь графики легко создаются с помощью SPSS, Microsoft Excel и других программ. Круговая диаграмма является отличным способом отображения относительных размеров категорий переменной. На Рисунке 3.1 показано количество человек получающие образование Республике Казахстан в 2018 г.

Пирог Чан делит круг на «кусочки», пропорциональные относительным частотам категорий. Самый большой срез представляет среднее образование, с почти 60%, а наименьший срез представляет профессиональное, как наименьшую категорию. Вы можете судить о круговой диаграмме как о визуальном частотном распределении, и на Рисунке 3.1 четко показаны относительные размеры типов образования.

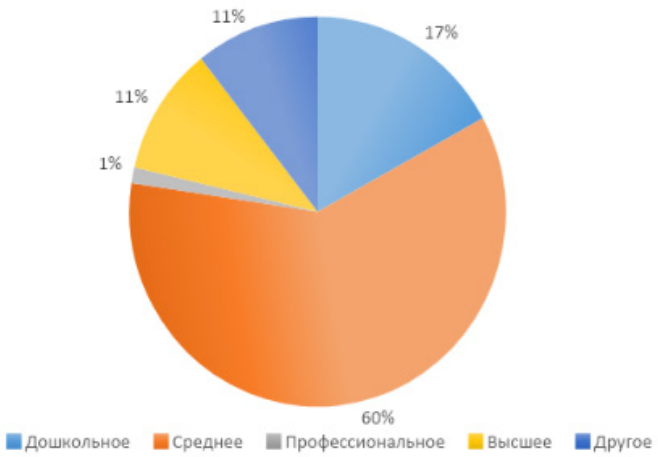

Рисунок 3.1 Количество человек получающие образование в Республике Казахстан, 2018 г.

Таблица 3.13. Количество человек получающие образование в Республике Казахстан, 2018 г.

Образование	Количество (чел)	%
Дошкольное образование	862 305	17
Среднее образование	3 050 800	60
Профессиональное образование	64 814	1
Высшее образование	542 458	11
Другое	537 917	11
Итого	5 058 294	100

Источник. См. таблицу 3.2.

Как и круговые, гистограммы являются прямыми. Категории переменной располагаются вдоль горизонтальной оси (или абсцисс), а частоты или проценты, если вы предпочитаете, вдоль вертикальной оси (или ординаты). Высота столбцов пропорциональна относительным частотам категорий и, как и круговые диаграммы, гистограммы являются визуальными эквивалентами частотных распределений. Рисунок 3.2 будет интерпретирован точно так же, как и круговая

диаграмма на рисунке 3.1, и исследователи могут свободно выбирать между этими двумя методами отображения данных. Однако, если переменная имеет более четырех или пяти категорий, предпочтительной будет гистограмма, так как круговая диаграмма переполнена большим количеством категорий.

Рисунок 3.2. Количество человек получающие образование в Республике Казахстан, 2018 г.

Линейные диаграммы (или частотные полигоны) аналогичны гистограммам, но связаны с линиями и расположены в средней точке интервалов, чтобы представить частоты. Высота точки отражает n случаев в интервале. Эти графики особенно подходят для переменных уровня со многими оценками.
Гистограммы и линейные диаграммы являются альтернативными способами отображения одного и того же сообщения. Таким образом, выбор между этими двумя методиками остается на верхушке тематических удовольствий исследователя.

Подведем итоги:

1. Мы рассмотрели плотности распределения - показатели, которые обобщают все распределения некоторой переменной. Статистический анализ почти всегда начинается с расчётов этих показателей для каждой переменной. Колонки для процентов, совокупных частот и/или совокупных процентов часто увеличивают восприятие плотностей распределения.
2. Проценты и пропорции, отношения, ставки дают нам различные способы выразить наши результаты с точки зрения относительной частоты. Проценты и пропорции сообщают об относительном возникновении некоторой категории переменной по сравнению с распределением. Отношения сравнивают два категории друг с другом, и ставки сообщают о фактических случаях некоторого явления по сравнению с количеством возможных случаев за некоторую единицу времени. Процентное изменение показывает относительное увеличение или уменьшение в переменной со временем.
3. Круговые диаграммы и гистограммы, гистограммы и линейные диаграммы (или многоугольники частоты) являются графиками, которые выражают информацию, содержавшуюся в плотности распределения компактным и визуально драматическим способом.