# Выборка

Методы отбора, расчета и анализа

# Термины

- Генеральная совокупность
- Выборка
- Доверительный интервал
- Процедуры анализа

# Выборка

- Выборка – часть генеральной совокупности, соответствующая ей по заданным критериям (характеристикам)

- Невероятностный метод построения
  1. Выборка согласных
  2. Выборка по усмотрению
  3. Метод квот
  4. Метод снежного кома (snowball)

- Вероятностный метод – классификация способов
  - Отбор элементов или кластеров
  - Все единицы отбора имеют равную вероятность для включения в выборку
  - Применяется или нет стратификация
  - Простой или систематический случайный отбор
  - Отбор через одну или несколько стадий

# Типы выборки

| Тип выборки | Описание |
|---|---|
| Простая случайная | Лотерейный выбор из генеральной совокупности |
| Стратифицированная случайная | Совокупность делится на страты (сегменты), внутри выборка случайная |
| Районированная (кластерная) | Случайный выбор групп (районов), внутри которого опрашивается каждый |
| Многоступенчатая | Случайно выбранные группы – города – улицы - дома |
| Доступная | Доступные в данное время респонденты |
| Квотированная | Выбираются квоты и опрашиваются респонденты |

# Описание выборки

- Исследуемая совокупность …
- Основа выборки
- Способ построения выборки
- Размер выборки
- Процесс построения выборки

# Определение размера выборки

- Произвольный (напр. 5% от совокупности)
- По аналогичным исследованиям
- Стоимость исследования
- Расчетный метод – мин. объем с т.зр. надежности и достоверности
- На основании доверительного интервала

Выборка или перепись?

- Ошибка выборки – точность выборки

# Определение размера выборки расчетным методом

$$n = \frac{Z^2(pq)}{e^2} \qquad n = \left(\frac{Z\sqrt{p(1-p)}}{e}\right)^2$$

- где n – объем выборки;
- z – нормированное отклонение, определяемое исходя из выбранного уровня доверительности - a;
- p – найденная вариация для выборки;
- q = (100 – p);
- e – доверительный интервал, в десятичной форме, желаемая погрешность (например, 0,04 = ±4%).

Значение нормированного отклонения оценки z от среднего значения в зависимости от доверительной вероятности (a) полученного результата

| **a, %** | 60 | 70 | 80 | 90 | 90 | 95 | 97 | 99,0 | 99,7 |
|---|---|---|---|---|---|---|---|---|---|
| **z** | 0,84 | 1,03 | 1,29 | 1,44 | 1,65 | 1,96 | 2,18 | 2,58 | 3,0 |

# Доверительный интервал

- Невозможно узнать истинное значение среднего генеральной совокупности на основе данных выборки
  - Ошибка выборки неизбежна
- Но можно оценить интервал значений, в который с определённой вероятностью входит истинное значение среднего
- Такой интервал называется **доверительным интервалом**
- Для нахождения доверительного интервала, мы сначала определяем вероятность, с которой мы хотим быть уверены в нашей оценке истинного среднего значения
  - как правило мы хотим быть уверены в нашей оценке как минимум на 95% или хотим доверять нашей оценке истинного значения среднего на 95%.

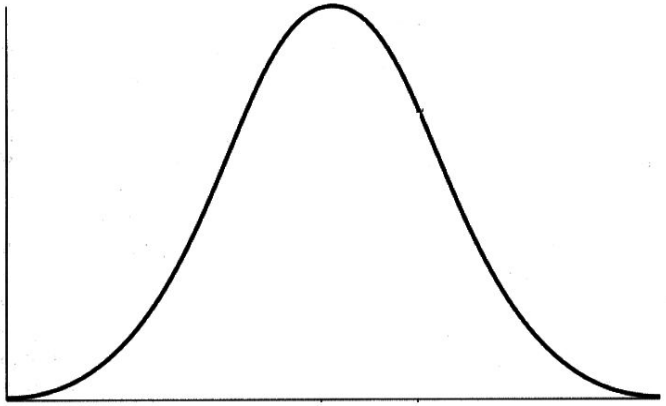
# Вероятность

Вероятность – это мера возможности появления события (благоприятного исхода)

$$p = \frac{\text{Число благоприятных исходов}}{\text{Общее число исходов}}$$

# Распределение вероятностей

- Имеет среднее, дисперсию и стандартное отклонение, которые помечаются следующими буквами:
- μ (мю) –среднее распределения вероятности
- $\sigma^2$ и σ (сигма) дисперсия и стандартное отклонение
- Форма распределения – Гауссова кривая (нормальное распределение



Площадь под кривой равняется 100 % всех наблюдений или вероятности = 1,0

# Частотное распределение и распределение вероятности

- Распределение вероятности (РВ) основано на теории вероятности, а частотное распределение (ЧР) основано на эмпирических (наблюдаемых) данных
- РВ – идеал, ЧР - реальность
- у РВ форма нормального распределения, а у ЧР форма приближается к нормальному распределению наблюдений

# Как перевести любое значение переменной в стандартную оценку (z-score)?

μ = среднее

σ = стандартное отклонение

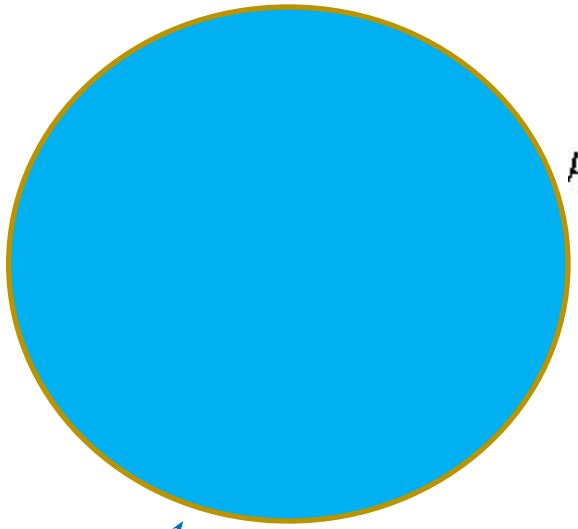
*z* = стандартная оценка

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

Стандартная оценка говорит нам, на сколько стандартных отклонений выше или ниже данное значение переменной от среднего значения

Используя эту информацию можно рассчитать вероятность того, что переменная принимает значения выше или ниже заданного  (заданных) показателей

# Стандартное отклонение среднего




$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

# Доверительный интервал

- Затем мы рассчитываем расстояние от средней, которое охватит 95% распределения выборочных средних по формуле:

$$\textbf{95\% доверительный интервал} = \bar{X} \pm 1.96\,\sigma_{\bar{X}}$$

- Мы используем 1,96, потому что это стандартная оценка (z-оценка), которая обозначает границы в 95% от всей площади под кривой распределения вероятности
- И, наконец,  устанавливаем границы интервала
- Полученный интервал указывает нам минимальное и максимальное значения переменной между, которыми мы можем утверждать с уверенностью в 95% находится истинное значение среднего (т.е. среднее в генеральной совокупности)

# Доверительный интервал

- **Доверительный интервал**, можно понимать как погрешность, задает размах части кривой распределения по обе стороны от выбранной точки, куда могут попадать ответы.
- Например, выборка в 384 человека для генеральной совокупности более 500 000 человек (например, один из районов города) означают доверительную вероятность 95% и доверительный интервал ±5%. То есть при проведении 100 исследований с такой выборкой (384 человека) в 95 процентов случаев получаемые ответы по законам статистики будут находиться в пределах ±5% от исходного.

# Доверительная вероятность

- показывает, с какой вероятностью случайный ответ попадет в доверительный интервал. Для простоты можно понимать её как точность выборки. Как правило, используется 95%, но при низких бюджетах ее можно уменьшить до 90% или 85%.Это приведет к снижению точности, что нужно учесть в выводах.

# Расчет выборки -2

- Если мы знаем размер генеральной совокупности:

1) n – объем выборки;
2) N - размер генеральной совокупности:
3) z – нормированное отклонение, определяемое исходя из выбранного уровня доверительности;
4) p – найденная вариация для выборки;

q = (100 – p);

5) $\Delta$ – допустимая ошибка – 5%.

$$n = \frac{z^2 pqN}{\Delta^2 N + z^2 pq}$$

## Пример. Таблица определения объема выборки

Правила вычисления

Найдите в верхнем ряду соответствующий объем выборки (1)

Найдите в колонке расчетную долю генеральной совокупности (2)

В месте пересечения ряда и колонки указана степень точности (± проценты).

| (1) ⇨ (2) | 100 | 200 | 300 | 400 | 500 | 600 | 800 | 1000 | 1200 | 1500 | 2000 | 2500 | 3000 | 4000 | 5000 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 % или 95% | 4,4 | 3,1 | 2,5 | 2,2 | 2,0 | 1,8 | 1,5 | 1,4 | 1,3 | 1,1 | 0,96 | 0,87 | 0,79 | 0,69 | 0,62 |
| 10% или 90% | 6,0 | 4,3 | 3,5 | 3,0 | 2,7 | 2,5 | 2,1 | 1,9 | 1,7 | 1,6 | 1,3 | 1,2 | 1,1 | 0,95 | 0,85 |
| 15% или 85% | 7,1 | 5,1 | 4,1 | 3,6 | 3,2 | 2,9 | 2,5 | 2,3 | 2,1 | 1,9 | 1,6 | 1,4 | 1,3 | 1,1 | 1,0 |
| 20% или 80% | 8,0 | 5,7 | 4,6 | 4,0 | 3,6 | 3,3 | 2,8 | 2,5 | 2,3 | 2,1 | 1,8 | 1,6 | 1,4 | 1,3 | 1,1 |
| 25% или 75% | 8,7 | 6,1 | 5,0 | 4,3 | 3,9 | 3,6 | 3,0 | 2,8 | 2,5 | 2,3 | 1,9 | 1,7 | 1,6 | 1,4 | 1,2 |
| 30% или 70% | 9,2 | 6,5 | 5,3 | 4,6 | 4,1 | 3,8 | 3,2 | 2,9 | 2,7 | 2,4 | 2,0 | 1,8 | 1,7 | 1,4 | 1,3 |
| 35% или 65% | 9,5 | 6,8 | 5,5 | 4,8 | 4,3 | 3,9 | 3,3 | 3,1 | 2,8 | 2,5 | 2,1 | 1,9 | 1,7 | 1,5 | 1,4 |
| 40% или 60% | 9,8 | 7,0 | 5,7 | 4,9 | 4,4 | 4,0 | 3,4 | 3,1 | 2,8 | 2,5 | 2,2 | 2,0 | 1,8 | 1,5 | 1,4 |
| 45% или 55% | 9,9 | 7,0 | 5,8 | 5,0 | 4,5 | 4,1 | 3,5 | 3,2 | 2,9 | 2,6 | 2,2 | 2,0 | 1,8 | 1,6 | 1,4 |
| 50% | 10,0 | 7,1 | 5,8 | 5,0 | 4,5 | 4,1 | 3,5 | 3,2 | 2,9 | 2,6 | 2,2 | 2,0 | 1,8 | 1,6 | 1,4 |

Эта таблица поможет определить степень точности при каждом значении расчетной доли генеральной совокупности для заданного объема выборки, или объем выборки при заданной степени точности. Если объем выборки равен 500 респондентам, то степень точности для предполагаемого значения доли генеральной совокупности, равного 5% (или 95%), составит ±2%, а для доли, примерно равной 20% (или 80%), составит ±3,6%. Иначе говоря, если исследователь хочет определить необходимый объем выборки, чтобы оценить результаты со степенью точности ±5% при значении доли около 50%, ответом будет 400 респондентов.