Evaluation: Statistics

Vojtěch Horký Peter Libič Petr Tůma

2010 – 2021

Contents

1 Overview

About

What to do with measurements once collected ?

Exploration

Examine measurements to identify basic properties and decide future analysis or future measurements.

Filtering

Removing observations that are not relevant for further analysis.

Summarization

Computing summary statistics about observations.

Formulating Conclusions

Testing whether observations support particular conclusions.

... usually iterative

Can We Use Statistics ?

Measurement data often fluctuates:

- System noise and measurement accuracy
- Impact of system observation
- Changes in system workload

Good Fit !

Statistics derives information about system from limited observations.

Really ?

Ever seen a statistics paper begin with "Let $X_1, X_2 ... X_n$ be i.i.d. random variables" ?

- Random
- Independent
- Identically distributed

Unfortunately, these assumptions (almost) never hold.

Statistical Distributions

Distribution Characterization

Cumulative Distribution Function (CDF)

$$F_X(x) = P(X \leq x)$$

... the one that looks like a snake

Probability Density Function (PDF)

$$f(x) = \frac{dF(x)}{dx}$$

... the one that looks like a camel

Common Distribution Parameters

Expected Value (Mean)

$$\mu = E(X) = \sum_{i=1}^{n} p_i x_i$$

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

... intuitively understood as representative value

Quantiles

$$P(X \leq x_\alpha) = F(x_\alpha) = \alpha$$

... value of x where CDF reaches α

Common Distributions in Performance Evaluation

In performance evaluation and modeling, it is useful to know about these distributions:

- Uniform distribution
- Normal distribution
- Exponential distribution
- (Binomial distribution)

What For ?

Mostly modeling. Sometimes approximation. Do not expect measurements to fit any analytical distribution !

Uniform Distribution

All values from the given range have the same probability.

Notation

$$U(a, b)$$

Density

$$f(x) = \frac{1}{b-a}, \text{for } a \leq x \leq b$$

Mean

$$E(X) = \frac{a+b}{2}$$

Median

$$\frac{a+b}{2}$$

Variance

$$V(X) = \frac{(b-a)^2}{12}$$

Exponential Distribution

Time between events that appear independently at constant average rate.

Notation

$$Exp(\lambda)$$

Density

$$f(x, \lambda) = \lambda e^{-\lambda x}, \text{for } x \geq 0$$

Mean

$$E(X) = \frac{1}{\lambda}$$

Median

$$\frac{ln(2)}{\lambda}$$

Variance

$$V(X) = \frac{1}{\lambda^2}$$

The distribution is memoryless, past waiting does not change future probabilities.

Normal Distribution

Value that sums many independent components.

Notation

$$N(\mu, \sigma^2)$$

Density

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

Mean

$$E(X) = \mu$$

Median

$$\mu$$

Variance

$$V(X) = \sigma^2$$

Binomial Distribution

Success count in coin toss experiments.

Notation

$$B(n, p)$$

Mass

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

Mean

$$E(X) = np$$

Median

$$\text{between} \lfloor np \rfloor \text{ and } \lceil np \rceil$$

Variance

$$V(X) = np(1-p)$$

For large n we can approximate $B(n,p) \approx N(np, np(1-p))$.

Central Limit Theorem

Let $X_1, X_2, \ldots$ be independent identically distributed (*iid*) random variables, $E[X_i] = \mu$, $\sigma^2 < \infty$. Then

$$\sqrt{n}((\frac{1}{n}\sum_{i=1}^{n} X_i) - \mu) \text{ converge in distribution to } N(0, \sigma^2)$$

Practically ?

Averages (and sums) of *enough* statistically *independent* samples will have (close to) normal distribution.

Practical pitfalls:

- How many summands are enough ? Some people will tell you a number, like 30 ...
- How to guarantee independency ?

Central Limit Theorem

Play a bit with the Central Limit Theorem. Following are some examples in R:

```
library ('tidyverse')

SUMS <- c (1, 2, 3, 5, 10, 20, 50, 100)
SAMPLES <- 1000

show_sums <- function (data) {
    samples <- bind_rows (
        map (SUMS, function (sum_size)
            tibble (
                size = sum_size,
                sample = replicate (SAMPLES,
                    sum (data [sample.int (length (data), sum_size, replace = TRUE)])
                )
            )
        )
    )

    ggplot (samples) +
        geom_histogram (aes (x = sample), bins = 33) +
        facet_wrap (facets = vars (size), scales = 'free')
}

show_sums (rnorm (SAMPLES))
show_sums (runif (SAMPLES))
show_sums (rexp (SAMPLES))
```

A bit more involved example introduces dependency between samples:

```
library ('tidyverse')

SUMS <- c (1, 2, 3, 5, 10, 20, 50, 100)
SAMPLES <- 1000
DEPENDENCY <- 0.9

sample_with_dependency <- function (data, sample_size) {
    sample <- data [sample.int (length (data), sample_size + 1)]
    for (index in seq_len (sample_size)) {
        sample [index+1] <- sample [index] * DEPENDENCY + sample [index+1] * (1 - DEPENDENCY)
    }
    return (sample [seq_len (sample_size)])
}
```

```
show_sums_dependent <- function (data) {
    samples <- bind_rows (
        map (SUMS, function (sum_size)
            tibble (
                size = sum_size,
                sample = replicate (SAMPLES,
                    sum (sample_with_dependency (data, sum_size))
                )
            )
        )
    )

    ggplot (samples) +
        geom_histogram (aes (x = sample), bins = 33) +
        facet_wrap (facets = vars (size), scales = 'free')
}

show_sums_dependent (rnorm (SAMPLES))
show_sums_dependent (runif (SAMPLES))
show_sums_dependent (rexp (SAMPLES))
```

Measurements ?

Random Variables

Measurements considered as *realizations* $x_1, x_2 ... x_n$ of *random variables* $X_1, X_2 ... X_n$.

- The interesting information is in the population properties of X_i
- We only have the sample properties observed on x_i
- We want to estimate the former from the latter

We must be aware that the model is not entirely accurate:

- It is especially difficult to get independent observations
- We are balancing model complexity with accuracy
- Guarantees are hard to get

We construct empirical distribution function approximations:

- Sample frequencies used instead of probabilities
- Various techniques for smoothing the result

Measurements ?

We can view measurements as random samples used to infer unknown constant parameter. Issues:

- The limitations (deficiencies) of the experiment
- Influence of the environment:
  - Initial system state is not identical for each experiment
  - Some state transitions non deterministic (external influence)

We refer to variability as measurement *error* or *noise* (more common).

Systematic Errors

Introduce bias into observations. Must be addressed by experiment design.

Random Errors

Nondeterministic, unpredictable, sometimes partially controllable. Long term impact averages out to zero by definition (unbiased).

Systematic Errors

Minimizing

- Use accurate measurement techniques.
- Minimize external interference.

But be careful to retain realistic variability !

Converting

Sometimes systematic error can be converted into random error:

- Randomize memory layout
- Randomize experiment order
- Randomize repetition count
- ...

Random Errors

We can toy with statistical interpretation:

- Assume we have many independent error sources.
- Assume that each error source:
  - adds $+E$ to the measured value with probability of $1/2$
  - adds $-E$ to the measured value with probability of $1/2$

For n error sources, the range of measurements is thus $(x - nE, x + nE)$ for the unknown true value x.

Furthermore:

- Each source of error either adds or subtracts E.
- Total error determined by the (random) number of sources that add E.
- This can be interpreted as trials from *binomial distribution*.
- For large n approximated by *normal distribution*.

... or we can wave hands and invoke CLT

But ...

... why, then, do we observe mostly asymmetric long tail noise ?

Many possible reasons:

- Error sources dependent
  - External interrupt means more instructions executed
  - But also caches trashed so more cache misses
  - And that may imbalance thread scheduling
  - And that may add to passive waiting
  - And ...
- Disruption likely to impact consecutive samples
- Error sources of (very) different magnitudes
- ...

2 Filtering

Why Filter Observations ?

Irrelevant Observations

We may have collected observations irrelevant to our investigation:

- Initial measurements taken during warm up
- Measurements distorted by external factors
- Measurements whose context is not relevant
- ...

Disruptive Observations

The observations may have properties that disrupt processing:

- Outliers that cause plot scale to stretch
- Outliers that inflate variance estimates
- Outliers that shift mean estimate
- …

Warm Up

System execution phase where performance is influenced by one time artifacts that are not relevant to long term performance.

Detecting From Measurements

Observe measurement properties:

- Wait for end of decreasing trend
- Wait for segment of stable observations
- Wait for new measurements to bring no new information
- …

Do not do this if you can help it.

Detecting From System State

Observe additional relevant system state:

- Wait for compilation events to cease for some time
- Wait for cache miss rate to stabilize
- …

Outliers

Typically identified using distance from some aggregated metrics:

- Values far away from mean
- Values outside certain quantiles
- Values far away from nearest neighbor

Thresholds typically relative:

- Further away than $3 \times \sigma$
- Further away than inter quartile range

Applied on all data or on sliding window.

Winsorization

Rather than discarding outliers altogether, replace them with the nearest value that is not considered an outlier.

3 Summarization

About Aggregate Statistics

Information Loss

Single number always loses information. But it can help show summary properties.

Consider robustness:

- A single outlier observation can move the mean anywhere. Mean is said to have a breakdown point of 0%.
- Including at least half of outliers in observations will move the median anywhere. Median is said to have a breakdown point of 50%.
- Including at least N% of outliers in observations will move the N% trimmed mean anywhere. N% trimmed mean is said to have a breakdown point of N%.

But robust does not always mean stable.

[1]

[1] See also `https://www.autodesk.com/research/publications/same-stats-different-graphs`

3.1 Central Tendency

Describing Central Tendency

Measures of central tendency like mean or median give a single number that is somehow common or typical or representative of the observations.

Mean is the average value of observations. Useful with observations whose sum makes sense.

Median is the value in the middle of a sorted series of observations. Useful with observations whose distribution is asymmetric.

Mode is the most likely value of a series of observations. Useful with observations whose domain is discrete.

Midrange is the average value of minimum and maximum observation.

[2]

What Mean to Use?

Trimmed Mean

Trimmed mean is a mean calculated after values outside certain quantile range are replaced with the quantiles (or discarded). Usually, but not necessarily, arithmetic mean is used. Useful for situations where the presence of outliers needs to be reflected but the outlier values themselves make the standard mean useless. **Example:** Sequence 1-10, outlier 100: mean 14.1, 10% trimmed mean 6.

```
x = c(1:10, 100)
mean(x)
mean(x, trim = 0.1)
```

What Mean to Use?

Weighted Mean

Weighed mean is a mean calculated with values counted according to their weight. Usually, but not necessarily, arithmetic mean is used. Useful for situations where all observations do not have equal significance. **Example:** Sequence 1-10, weights 10-1: mean 5.5, weighted mean 4.

```
x = 1:10
w = 10:1
mean(x)
weighted.mean(x, w)
```

What Mean to Use?

Arithmetic Mean

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ Used for variables that are additive in nature (for example durations). **Example:** Operations take 1, 2 and 3 seconds: arithmetic mean is 2 seconds, sequence of 3 operations will take $2 \cdot 3 = 6$ seconds.

```
mean(x)
```

[2] See also `https://xkcd.com/2435`

What Mean to Use?

Geometric mean

$\dot{x} = \sqrt[n]{\prod_{i=1}^{n} x_i}$ Used for variables that are multiplicative in nature (for example ratios). **Example:** Optimizations speed up 10, 20 and 30 percent: geometric mean is 19.7 percent, sequence of 3 optimizations will speed up $1.197^3 = 1.716$ times or 71.6 percent.

```
x <- c(10, 20, 30)
exp(mean(log(x)))
```

What Mean to Use?

Harmonic mean

$\ddot{x} = \frac{n}{\sum_{i=1}^{n} 1/x_i}$ Used for variables whose inverse value is additive in nature (for example throughput). **Example:** Server performs first 1000 operations at 100 operations per second, the next 1000 operations at 200 operations per second: harmonic mean is 133 operations per second, sequence of 2000 operations will take $2000/133 = 15$ seconds.

```
x <- c(1000,100)
1/mean(1/x)
```

Median

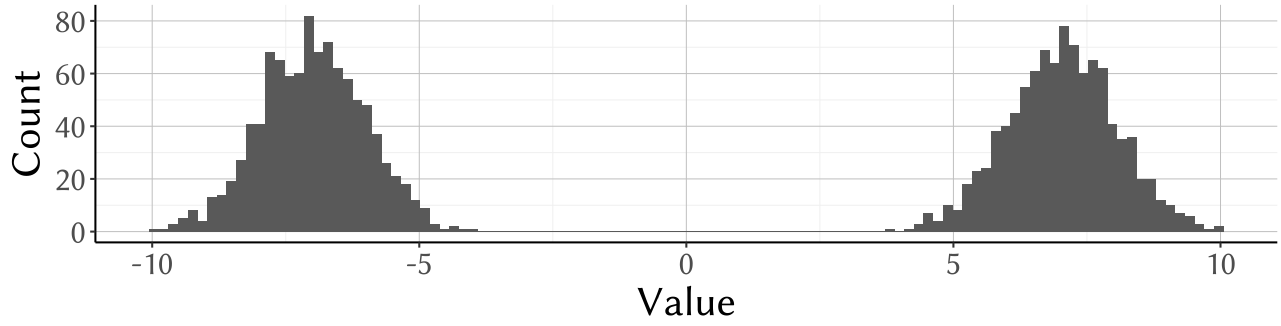
Median

$\tilde{x} = x_{(n+1)/2}$ for odd n $\tilde{x} = (x_{n/2} + x_{n/2+1})/2$ for even n

Useful with observations whose distribution is asymmetric. And when it makes sense to report one concrete value.

```
median(x)
```

Can you guess the median for the histogram below ?



Mode

Modes

Useful with observations whose domain is discrete.

```
as.numeric (names (which.max (table (x))))
```

But also possible to estimate for continuous domains.

```
d <- density (x)
d$x [which.max (d$y)]
```

Median, Mean and Mode in Different Distributions

Think about location of median, mean and mode in distributions that are:

- Unimodal symmetrical
- Right (positive) skewed unimodal
- Left (negative) skewed unimodal
- Bimodal symmetrical
- Uniform

Mean-Median-Mode Inequality

In many cases, $mean \leq median \leq mode$ or $mean \geq median \geq mode$. But this is mostly just fun fact ...

The following function can come in handy when playing with mean-median-mode inequality.

```
library ('tidyverse')

show_mmm <- function (x) {
    x_mean <- mean (x)
    x_median <- median (x)
    x_density <- density (x)
    x_mode <- x_density $ x [which.max (x_density $ y)]

    print (ggplot (tibble (x = x)) +
        geom_histogram (aes (x), bins = 111) +
        geom_vline (aes (xintercept = x_mean, color = 'mean')) +
        geom_vline (aes (xintercept = x_median, color = 'median')) +
        geom_vline (aes (xintercept = x_mode, color = 'mode')) +
        scale_color_manual (values = c ('mean' = 'red', 'median' = 'green', 'mode' = 'blue')))
}
```

3.2 Variability

Describing Variability

Range gives the minimum and maximum observations.

Variance gives a measure of variability.

$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Useful with observations whose distribution is close to normal or at least symmetric.

Standard deviation is a square root of variance, $\sigma = \sqrt{\sigma^2}$. Units of standard deviation are the same as units of observations.

Three sigma rule

For intuitive understanding, three sigma rule says it is rare to see observations more than three sigma away from the mean. The actual percentages are 68.3% - 95.5% - 99.7%, only about 0.1% to each side of three sigma, for normal distribution.

Coefficient of variation is the ratio of standard deviation to mean, $\sigma/\bar{x}$. Is dimensionless and therefore tempting, but only works for ratio scales.

Mean absolute deviation gives a measure of variability from the mean.

$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

Median absolute deviation gives a measure of variability from the median.

$median_i(|x_i - median(x)|)$

Quantile range gives the range between selected quantiles.

Inter-percentile range is between 10% and 90% of observations.
Inter-quartile range is between 25% and 75% of observations.

Interval and Ratio Scale

A scale is denoted as *interval* scale when computing a difference between two values makes sense, but computing a ratio of two values does not. Temperature scales (except for those with zero at absolute zero) are interval scales. A scale is denoted as *ratio* scale when both computing a difference and computing a ratio makes sense.

Describing Variability

Range `min(x); max(x)`
Variance `var(x)`
Standard deviation `sd(x)`
Coefficient of variation `sd(x)/mean(x)`, not `cov(x)`!
Mean absolute deviation `mean(abs(x - mean(x)))`
Median absolute deviation `mad(x, constant=1)`
Inter-percentile range `quantile(x, percents)`
Inter-quartile range `IQR(x)`

Sample Average Properties

$$E(X_i) = \mu$$

$$E(\overline{X_n}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) \overset{X_i \text{iid}}{=} \mu$$

$$var(X_i) = \sigma^2$$

$$var(\overline{X_n}) = var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) \overset{X_i \text{iid}}{=} \frac{\sigma^2}{n}$$

$$sd(X_i) = \sigma$$

$$sd(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}$$

Confidence Intervals

We are estimating *population* parameter from *sample* parameters. Can we express how certain we are ?

Confidence Interval

A confidence interval for parameter C at confidence level $\gamma = 1 - \alpha$ is an interval (C_{lo}, C_{hi}) such that *across many experiments*, $P(C \in (C_{lo}, C_{hi})) = \gamma$.

For correct confidence interval sometimes also $P(C \leq C_{lo}) = \alpha/2$ $P(C \geq C_{hi}) = \alpha/2$.

Confidence level:

- Higher confidence means larger interval
- Commonly used are 95% and 99%
- Beware correct interpretation

Confidence Intervals

The confidence level γ concerns the *procedure* but not the *individual intervals*.

What it *does* mean

If we keep making new sets of observations and constructing confidence intervals from them, a share of γ intervals will contain the unknown true mean.

What it *does not* mean

If we compute single confidence interval, the probability that it contains the unknown true mean is γ.

Why so, and is this such a big difference ?

Confidence Intervals

Measurement

Consider tossing a fair coin and reporting either X or $X + 1$ with equal probability.

Interval Construction Procedure

Make two observations, these are either (X, X), $(X, X + 1)$, $(X + 1, X)$, or $(X + 1, X + 1)$. Given (o_1, o_2), our confidence interval for unknown X will be computed as $< min(o_1, o_2), max(o_1, o_2) >$.

- Across many trials our interval will contain X in 75% cases. This makes it an interval for X with confidence level 75%.
- But assume a single trial:
  - Either $o_1 = o_2$ and then we cover X with 50% chance,
  - or $o_1 \neq o_2$ and then we are certain to cover X.

[3]

Confidence Interval Controversy

There is actually a long running discussion about correct confidence interval interpretation. Check out doi:https://dx.doi.org/10.3758/s13423-013-0572-3 and doi:https://dx.doi.org/10.3758/s13423-015-0955-8 for more.

Computing Confidence Intervals

$$\overline{X_n} \sim N(\mu, \frac{\sigma^2}{n}) \qquad\qquad Z = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

For $n \geq 30$ (typically) we can approximate σ with s:

$$C_{lo} = \overline{X_n} - z_{1-\alpha/2}\frac{s}{\sqrt{n}} \qquad\qquad C_{hi} = \overline{X_n} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

$$P(Z \leq z_{\alpha/2}) = \alpha/2$$

For lower n, we should use $t_{\alpha/2;n-1}$ instead of $z_{\alpha/2}$

```
# The hard way ...
c_lo = mean(x) - qt (1 - alpha/2, n - 1) * sd(x) / sqrt (n)
c_hi = mean(x) + qt (1 - alpha/2, n - 1) * sd(x) / sqrt (n)

# The easy way ...
t.test (x, conf.level = gamma)$conf.int [1:2]
```

[3] Based on MacKay: Information Theory ... http://www.inference.org.uk/mackay/itila

How Many Measurements ?

How many measurements are needed for specified confidence level $\gamma = 1 - \alpha$ and (relative) width of interval ϵ ?

If

$$CI = (1 \pm \epsilon)\overline{X_n} = \overline{X_n} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

then

$$n = (z_{1-\alpha/2}\frac{s}{\epsilon\overline{X_n}})^2$$

To apply this formula we need some $\overline{X_n}$ and s to start with, usually constant number of initial measurements used for this.

Non Normal Distribution ?

Logarithmic Transformation

When the errors are close to Log Normal Distribution, their logarithm should be closer to normal.

Batch Means

Rather than taking individual samples, take small batches and work with an average from each batch.

... or use bootstrap !

Bootstrap Intuition

Assume we have enough samples to characterize the population. What will happen if we draw a random sample from our samples ?

Resampling

For population F and parameter θ, collect a representative sample $\hat{F}$. From that sample, draw repeated samples (resamples) $\hat{F}^*$ and use the sampling distribution of $\hat{\theta}^*$ as a plug in for the distribution of $\hat{\theta}$.

Illustration in Hesterberg: What Teachers Should Know ... `https://arxiv.org/abs/1411.5279`

- Figure 4 for sampling from population
- Figure 5 for resampling from original population sample

Percentile Bootstrap Confidence Interval

Example interval for mean:

1. Collect (measure) sample set $\hat{X}$
2. Draw random sample $\hat{X}^*$ from $\hat{X}$ with replacement
3. Compute mean $\hat{\mu}^*$ of $\hat{X}^*$
4. Repeat from step 2 to create set of estimates $\{\hat{\mu}^*\}$
5. Use quantiles of $\{\hat{\mu}^*\}$ as confidence interval bounds

This is just for illustration, there are better procedures available !

```
library ("boot")
data <- rnorm (1000)
boot_data <- boot (data, function(x, i) mean (x[i]), R = 10000)
boot.ci (boot_data, type="perc", conf = 0.99)
```

Bootstrap Gotchas

Approximates Observed Parameter
Samples of $\hat{\theta}^*$ taken from $\hat{F}$ will approximate $\hat{\theta}$, not θ.

Sample Size Matters
Bootstrap from small sample $\hat{F}$ will underestimate variance, similar to difference between population and sample variance. Also sampling distribution will be coarse.

Skewness
Skewed distribution introduces asymmetry into confidence intervals.

Bias
Depending on circumstances bootstrap may introduce (additional) bias.

[4]

Comparing Confidence Intervals

Sometimes used to detect significant difference. Three possible outcomes:

Intervals do not overlap
Interpreted as (statistically) significant difference between alternatives.

Intervals overlap but means not within the other interval
Interpreted as outcome without clear conclusion.

Intervals overlap and means are within the other interval
Interpreted as no difference between alternatives.

Pros
- Easy to visualize
- Easy to implement
- (Seemingly) easy to interpret

Cons
- Quality of decision not really obvious *It is not the confidence level of the intervals !*
- With conservative (wide) intervals result often partial overlap
- With aggressive (narrow) intervals wrong conclusion highly likely

Confidence Intervals for Difference

Rather than checking two confidence intervals of means for overlap, we can construct a single confidence interval for difference of means.

Two important outcomes:

Interval does not straddle zero
Conclude that the two means are different at given confidence level.

Interval straddles zero
Interpreted as no evidence that the two means are different.

[4] Based on Hesterberg: What Teachers Should Know ... `https://arxiv.org/abs/1411.5279`

4 Formulating Conclusions

Hypothesis Testing

Null Hypothesis (H_0)
The default result of the test, what we remain with when we see no evidence to the contrary. Typically:

- Samples were drawn from two populations with equal means
- A parameter has no impact on the population mean
- ...

Alternative Hypothesis
The experiment hypothesis we are interested in. Exclusive alternative to the null hypothesis but not necessarily its logical negation.

Hypothesis testing checks whether we have a good enough reason to reject the null hypothesis in favor of the alternative hypothesis.

Test Statistic
A function of collected observations whose distribution would differ depending on whether the null hypothesis or the alternative hypothesis holds.

p-Value
Expresses likelihood of (at most) specific test statistic value assuming the null hypothesis holds.

The test rejects the null hypothesis if the p-value is below chosen threshold α, called *significance level.*

In other words, we reject the null hypothesis if it appears unlikely.

Hypothesis Testing Example

Imagine testing whether a coin is fair. In an experiment, there were 2 heads in 10 tosses.

Hypotheses
Null hypothesis is that the coin is fair. Alternative hypothesis is the opposite.

Test Statistic
We use the number of heads as the test statistic. Under the null hypothesis it would have a binomial distribution $B(10, 1/2)$.

The likelihood of having at most 2 heads in 10 tosses under the null hypothesis from binomial CDF is 0.0547.

At significance level 5 %, our experiment did not give us a strong enough reason to reject the null hypothesis.

Common Tests

t-Test
Assuming two populations with normal distributions (of equal variance), tests whether the means of the populations are equal.

Mann-Whiteny-Wilcoxon U-Test
Assuming ordinal (comparable) samples from two populations, tests whether it is equally likely for an observation from one population to be larger than one from the other and vice versa.

Kolmogorov-Smirnov Test
Assuming ordinal (comparable) samples from two populations, tests whether there is a significant difference between their empirical distribution functions.

... there are *many* more !

Significance Level

Type I Error is rejecting H_0 when it is in fact true.
Type II Error is not rejecting H_0 when it is false.
Test significance level α is *conditional* probability of Type I Error.

Imagine diagnostic test run at $\alpha = 1\,\%$. The test diagnoses a rare condition (usually 1 in 1000). What is the false alarm rate ? Not $1\,\%$ ...

In a sample of 1000000 observations ... ... we will have $1000000/1000 = 1000$ true positives.
In the remaining 999000 observations ... ... we will make $999000 * 1\,\% = 9990$ false rejections.

Assuming all true positives are detected reliably, the false alarm rate is $9990/(1000 + 9990) = 91\,\%$!

> In other words, if the alternative hypothesis is unlikely, most rejects are flukes.

Statistical And Practical Significance

Statistically Significant
Different from null hypothesis in a way that makes test statistic unusual.

Practically Significant
Different from null hypothesis in a way that has practical consequences.

Imagine comparing performance of two software functions. One is $N(1\,000\,000\,\mathrm{s}, 1)$, the other is $N(1\,000\,001\,\mathrm{s}, 1)$.

Standard t-test at $5\,\%$ significance level will tend to recognize this difference with mere tens of samples.

... use (some) size of effect measures !

> The following somewhat ugly R code shows how often the difference is detected:
>
> ```
> library ('tidyverse')
> data <- bind_rows (map (seq.int (10, 50), function (n)
>     tibble (n = n, p = mean (replicate (1000,
>         t.test (rnorm (n, 1000000, 1), rnorm (n, 1000001, 1)) $ p.value < 0.05)))))
> ggplot (data) + geom_point (aes (x = n, y = p))
> ```
>
> The result is obviously due to the very low sample variance employed. Similar effects can be demonstrated in samples with outliers, where – depending on the test used – the test behavior can change significantly.

Test Power

Type I Error is rejecting H_0 when it is in fact true.
Type II Error is not rejecting H_0 when it is false.
Test power $1 - \beta$ is *conditional* probability of *not making* Type II Error.

Some (more or less) obvious observations:

- Test power changes with test significance level. The higher the significance level, the lower the power.
- Test power usually depends on sample size.
- Test power changes with effect magnitude.

One way to tune test power is by choosing sample size so that effects of practical magnitude are detected reliably.

Repeated Testing

Hypothesis Fishing

Repeated testing of the same data with different hypotheses or test parameters.

Technical measures to prevent false positives inflation:

- Bonferroni Correction uses α/m when testing m hypotheses to achieve desired family wise error rate (FWER) α
- Benjamini-Hochberg Procedure adjusts α to achieve desired false discovery rate (FDR)
- ...

Procedural measures to prevent false positives inflation:

- Validating hypotheses with fresh data
- Mandatory experiment protocol registration

[5] [6]

The family wise error rate (FWER) is defined as the probability that at least one of the family of tests will exhibit a Type I Error.
The false discovery rate (FDR) is defined as the proportion of tests that rejected the null hypothesis due to a Type I Error.

Procedures such as repeated evaluation of stopping condition also constitute repeated testing ...

More Problems With Testing

In a binary quiz, there were 9 correct answers after 12 questions. At significance level 5 %, is this a result of random guessing ?

Binomial Distribution (Fixed Trial Count)

$B(n, p)$ is successes in n trials with success probability p in each trial.

Looking at CDF for $B(12, 1/2)$ says the chance of getting at least 9 correct answers in 12 questions is $p = 0.073$.

Negative Binomial Distribution (Fixed Failure Count)

$NB(n, p)$ is failures until n successes with success probability p in each trial.

Looking at CDF for $NB(3, 1/2)$ says the chance of getting at least as far as 12 questions is $p = 0.033$.

[7]

The point of this example is that the background story matters. Unfortunately, the background story is not necessarily visible from the data, which suggests that by adjusting the story, we can adjust the conclusions made.
Binomial distribution computation is `pbinom (8, 12, 1/2, lower.tail = FALSE)`. Negative binomial distribution computation is `pnbinom (8, 3, 1/2, lower.tail = FALSE)`.

Publication Pressure Bias

It seems natural to prefer publication of new discoveries to publication of experiments that did not show anything surprising ...

Effect Magnification

Assume each measurement consists of effect and error. Which measurement is more likely to make it to publication ?

- One where the error is negative and effect appears smaller
- One where the error is positive and effect appears bigger

Possible solutions:

[5] See also https://xkcd.com/882
[6] See also doi:10.1016/S1053-8119(09)71202-9
[7] Based on Reinhart: Statistics Done Wrong ... https://www.statisticsdonewrong.com

- Support for reproduction studies
- Artifact evaluation procedures
- ...

But the inherent conflict does not go away !

Regression To The Mean

Another demonstration of the same mechanism is regression to the mean. When experiment identifies extreme entities, the same entities will likely be more average in repeated experiment. This may give false impression of a trend.